6-30-2016

Bayesian Nonparametric Approaches To Multiple Testing, Density Estimation, And Supervised Learning

William Cipolli III
University of South Carolina

Bayesian Nonparametric Approaches to Multiple Testing, Density Estimation, and Supervised Learning

by

William Cipolli III

Bachelor of Science
Quinnipiac University 2011
Bachelor of Arts
Quinnipiac University 2011
Master of Science
University of South Carolina 2014

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosphy in

Statistics

College of Arts and Sciences

University of South Carolina

2016

Accepted by:

Timothy E. Hanson, Major Professor

Brian Habing, Committee Member

Joshua M. Tebbs, Committee Member

Alexander C. McLain, Committee Member

Lacy Ford, Senior Vice Provost and Dean of Graduate Studies

Acknowledgments

Firstly, and most importantly, I thank my advisor Tim Hanson. It is difficult to explain how enjoyable it has been to work under him. Even though we have completed research on challenging, rich problems - which I will admit sometimes took me longer to grasp than it should have - he was steadfast in ensuring I got through any brick walls I ran into. He fundamentally cares and helped push me in the direction of gathering exciting developments and presenting those results in a clear and honest manner. Tim's generosity, humor and patient guidance have been invaluable during my time at the University of South Carolina. Simply put, I could not have asked for a better PhD advisor.

Many of these statements are echoed for the other members of the Statistics department here at the University of South Carolina. I joined University of South Carolina because it seemed like a rigorous program that placed a strong value on community; I was not disappointed and I look forward to keeping in touch and coming back to visit often. I especially want to single out the three other members that make up my committee: Josh Tebbs, Brian Habing and Alex McLain.

I remember emailing Josh a few weeks before receiving my letter of acceptance and frequently thereafter as I decided to come here. Josh's advice then, and throughout my time in the department, has been paramount to my success; you can only hear that your notation is horrible so many times before you get it right. After taking two courses in my first year with Josh, I am happy he has continued to guide my scholarship; he has undeniably influenced my graduate career in an immensely positive way. I am proud, however, that I have returned the favor - I've taught him everything

he knows when it comes to squash!

Brian's humor belies his depth of knowledge and his ability to complete really interesting projects but makes interacting with him and sitting in his courses very enjoyable. It was in Brian's class that I first got the flavor of academic writing; what I learned in that course helped me write many pages of this dissertation. He has also endured many conversations about teaching and testing which have been remarkably productive for me and my teaching pedagogy.

Alex has provided me with invaluable knowledge, guidance and insight particularly on the second chapter of this dissertation. His ideas will likely lead to at least one more paper on the multiple testing subject and his writing tips were a big help in making this register at all on a scale of one to ten on readability.

Finally, I suppose I would be remiss if I did not thank the staff and owner of Cool Beans in Columbia for tolerating my often loud presence at least once a day (and sometimes much more than that) for the past four years; their coffee kept me going.

Looking back, I wish that I had spent time making the acknowledgments one hundred times longer and cutting out the rest of the dissertation so I could proportionately show my thanks and appreciation for the faculty, colleagues, students, friends and family that have supported me along the way. Thank you!

Abstract

This dissertation presents methods for several applications of Polya tree models. These novel nonparametric approaches to the problems of multiple testing, density estimation and supervised learning provide an alternative to other parametric and nonparametric models. In Chapter 2, the proposed approximate finite Polya tree multiple testing procedure is very successful in correctly classifying the observations with non-zero mean in a computationally efficient manner; this holds even when the non-zero means are simulated from a mean-zero distribution. Further, the model is capable of this for "interestingly different" observations in the cases where that is of interest. Chapter 3 proposes discrete, and smoothed approximate mixtures of Polya trees for application in mixed models and density estimation. Finally, Chapter 4 proposes a supervised learning procedure based on marginal, multivariate finite Polya trees. This approach is successful in correctly classifying observations in a variety of scenarios where the ten-fold cross validation was kept low. The proposed methodologies and applications show the versatility and flexibility of nonparametric Polya tree based methods and Chapter 5 outlines some obvious but rich extensions for future research.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The split between frequentists and Bayesians is, mainly, that the frequentists set the parameters of interest as fixed and Bayesians view parameters probabilistically with some prior information. Consider the task of calculating a confidence interval for a population parameter. Frequentists build their methodology and design their experiment based on the idea that the parameter of the model is fixed, taking on one and only one value, and to ensure a certain proportion, say .95, of repeated trials would yield a confidence interval that contains the true parameter's value.

Bayesians build their methodology and design their experiment based on the idea that the value of the parameter is fixed but whose plausible values can be reflected by some probability distribution defined by the prior information. Bayesian methodologies instead use the prior information and data to report the most likely values of the parameter given the two sources of information available. One of the reasons Bayesian approaches have become so popular among statisticians is because it allows us to use an expert's beliefs about the quantity of interest before modeling the data. If these beliefs are accurate, using the Bayesian approach allows us to amalgamate the data and these prior beliefs to create a model that uses the "full picture." In the absence of prior information Bayesian approaches are still useful as Bayesian models can fit certain models that are difficult or impossible for frequentists while using "noninformative" priors.

Both the Bayesian and frequentist approaches include parametric models. With parametric models we assume the data come from a particular underlying probability

distribution function, most infamously the Gaussian distribution. The parameters, i.e. mean and standard deviation, of the assumed probability distribution are generally of interest. If our assumptions about the underlying probability distribution are good the parametric approach works quite nicely but if the data deviate from that distribution greatly the parametric approaches may yield incorrect analyses.

Nonparametric approaches differ by not specifying the structure of the model but letting the data select the distribution from an infinite, highly flexible class, thus potentially mitigating the misspecification of the underlying distribution. These nonparametric methodologies allow statisticians to fit very flexible models by relaxing the usual parametric assumptions that are imposed in the usual parametric approaches, like the ubiquitous Gaussian assumption. Nonparametric tests, however, are not always the magic answer. Although nonparametric tests make less assumptions and are able to fit flexible models, nonparametric approaches are often less powerful and more difficult to interpret than their parametric counterparts, particularly when the parametric assumptions are very accurate. Although not always the best choice, nonparametric statistics offer us a solution in cases where the parametric assumptions fail.

Bayesian nonparametric approaches have increasingly attracted much theoretical and application research in the statistical and computer science fields due to exponential advances in computation which makes fitting these complex models feasible. These techniques, that enjoy the benefits of both the Bayesian and nonparametric approaches, are then utilized across many other disciplines to push forward solutions to many of the world's complex problems.

In the following dissertation proposal, the nonparametric Polya tree is used as a prior to create a flexible Bayesian approach to several existing statistical problems, including multiple testing, meta-analysis, GLMM, density estimation and classification. The flexibility provided by the nonparametric approach of the Polya tree allows

for methodologies that are robust to the usual Gaussian assumption while simultaneously being Bayesian which adds the ability of updating the posteriors of, and making inference about, key values in the model.

Chapter 2 offers a multiple testing procedure in which a discrete approximation to a Polya tree prior, centered at the usual Gaussian distribution, is proposed which enjoys fast, conjugate updating. The model is employed to analyze data from a mixture model. This new technique has been remarkably successful in correctly identifying units to be rejected in the multiple testing phase, as well as estimating the non-null distribution. These methods are demonstrated using extensive simulation and real data analysis accompanied by a Java web application.

Chapter 3 continues with the discrete approximation to a Polya tree prior and introduces two applications in nonparametric meta-analysis and random intercept Poisson regression, as well as a smoothed mixture of Polya trees that attempts to improve the usual mixture of Polya trees model that often suffers the disadvantages of yielding quite spiky density estimates. These methods are demonstrated using several real data analyses in which the results are compared to the usual mixture of Polya trees and Dirichlet process mixture models with some success.

Chapter 4 explores the problem of classification in which the nonparametric approach of the multivariate Polya tree realizes impressive results in simulations and real data analyses; performing similarly to or better than current approaches in many cases. The flexibility gained from eliminating certain distributional assumptions from the model can greatly improve the ability to correctly classify new observations, as even minor deviations from distributional assumptions could lead to missing an important feature in any one level's density. The proposed method is quite fast compared to other supervised classifiers and very simple to implement as there are no kernel tricks or initialization steps that greatly affect the model, like the kernel trick for SVM.

1.1 Dirichlet Process

1.1.1 Introduction to the Dirichlet Process

The Dirichlet distribution can be thought of as the multivariate generalization of the beta distribution and as the beta distribution is a conjugate prior for the binomial, the Dirichlet distribution is a conjugate prior for the multinomial distribution. Note that the beta distribution is the special case of a Dirichlet distribution for two dimensions.

Let $\mathbf{p} = (p_1, \ldots, p_k)'$ be the vector of probabilities for k different classes from a multinomial experiment where each $p_i \in (0, 1)$ and $||\mathbf{p}||_1 = 1$. Consider a Dirichlet prior on $\mathbf{p}$, i.e. $\mathbf{p} \sim DP(\alpha_1, \ldots, \alpha_k)$, where each $\alpha_i \in \mathbb{R}^+$ and $k \geq 2$. The Dirichlet process is conjugate, so given observations $\mathbf{x} = (x_1, \ldots, x_k)'$ the posterior is $\mathbf{p}|\mathbf{x} \sim DP(\alpha_1 + x_1, \ldots, \alpha_k + x_k)$. The Dirichlet distribution is a distribution of probabilities on a simplex, and therefore a distribution over probabilities for class random variables.

The Dirichlet process introduced by Blackwell (1973) is also considered a distribution over distributions. Let G be a Dirichlet process, i.e. $G \sim DP(\alpha, G_0)$ with base distribution G_0 and positive scaling parameter α. The Dirichlet process, G, is then a probability measure that has the same support as G_0.

There is a profound difference between G and G_0. G is a discrete distribution made up of countable point masses and G_0 may be any probability measure, continuous, discrete, or mixed. This difference has consequence, the fact that G is constructed to give discrete distributions with probability one (Ferguson, 1973) makes it unsuitable for general applications in nonparametrics, however it is well suited for the problem of placing priors on mixtures to provide an estimate of the particular number of mixture components as its relationship to the multinomial suggests.

1.1.2 Sampling from a Dirichlet Process

The issue of the Dirichlet process yielding discrete distributions with probability one is clear in the case of sampling. Random sampling from the Gaussian distribution yields unique samples, i.e. the probability that two samples are equal is zero. However since the Dirichlet process, G, is a discrete distribution this probability that two samples are equal is non-zero.

Sampling from a Dirichlet process, according to the "stick-breaking" construction of Sethuraman (1994), with $\alpha = \alpha_0$ and $G_0 = N(0, 10^2)$ can be completed using the following steps. Simulate random variables $\beta_1, \beta_2, \ldots$ from a beta distribution with shape parameters $a = 1$ and $b = \alpha_0$. Let $\mathbf{p}$ be an n-dimensional vector so that $p_1 = \beta_1$ and $p_i = \beta_i \prod_{j=1}^{i-1}(1 - \beta_j)$. Simulate random variables $\theta_1, \theta_2, \ldots$ from a Gaussian distribution with $\mu = 0$ and $\sigma^2 = 10^2$. Then $\mathbf{X}$ can be sampled discretely from $\theta_1, \theta_2, \ldots$ with probabilities $\beta_1, \beta_2, \ldots$ respectively.

Suppose G_0 is a Gaussian distribution, i.e. $G \sim DP(\alpha, N(0, 10^2))$. We can sample from the Dirichlet process which provides a discrete sample from the baseline distribution G_0. Figure 1.1 is a plot of the baseline G_0 along with one draw from a Dirichlet process with $\alpha = 2$, where the stick breaking was truncated to one hundred pieces.

1.1.3 Dirichlet Process Mixture Models and the Chinese Restaurant Process

Consider an empty Chinese restaurant with an unbounded number of tables. The first customer, X_1, enters the restaurant and sits down at a table, θ_1. Then a second customer X_2 enters the restaurant and sits down at the same table, θ_1, with probability $\alpha/(1+\alpha)$ or at a new table θ_2 with probability $1/(1+\alpha)$. This continues yielding the generalization for the n^{th} customer entering the restaurant who will sit at a new table with probability with probability $\alpha/(n+\alpha)$ and sits at already occupied table θ_k with probability $n_k/(1+\alpha)$ where n_k is the number of customers sitting at table

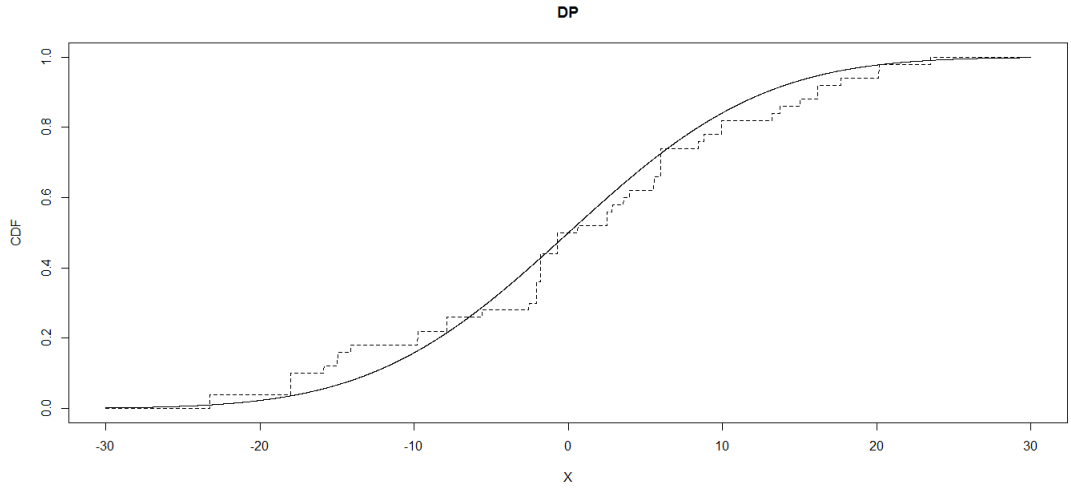

Figure 1.1: Dirichlet process sample from $N(0, 10^2)$ along with the Gaussian CDF.

θ_k.

The Chinese Restaurant Process (Aldous, 1985; Pitman, 2002), which gets its name from the metaphor above, has enjoyed much research and application in recent years as it has many desirable properties. The more data points there are at a cluster, the more likely it is that new data points will be assigned to that cluster but there's always a small probability that it can initiate its own, entirely new cluster. The choice of α is important here; a larger selection of α leads to more clusters and a smaller selection of α leads to less.

As seen in the Chinese Restaurant process, the Dirichlet process allows for an unspecified, and perhaps countably infinite, number of clusters. The same logic is extended to mixture models as explored by Antoniak (1974), Escobar and West (1995), and Rasmussen (1999). In traditional mixture modeling one must specify the number of clusters before analysis can begin; this Bayesian nonparametric approach, however, is flexible when it comes to this and uses the data to decide how many clusters should be required.

Recall that Dirichlet processes are constructed to give discrete distributions with probability one which is undesirable in many applications. The idea behind Dirichlet

process mixtures is to mitigate this concern by using a more complex model by convoluting G with a known kernel for smoothing. A Dirichlet process mixture does quite well in fitting the same $G_0 = N(0, 10^2)$ as seen in Figure 1.2 using just a random sample of $n = 50$ points drawn from G_0; even though the Dirichlet process is discrete, this Dirichlet process mixture model does provide a smooth estimate.

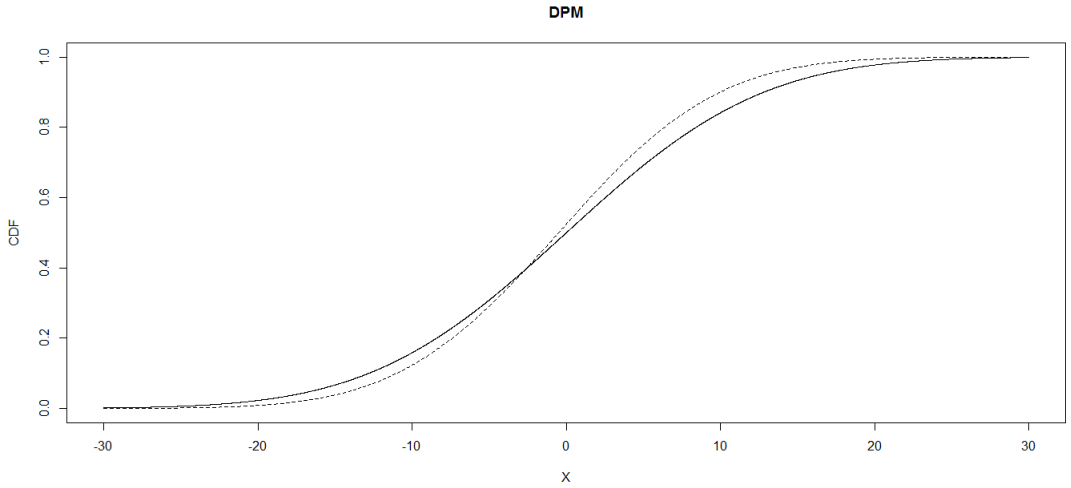

Figure 1.2: Dirichlet process draw from the $N(0, 10^2)$ cumulative distribution function.

This methodology has been very popular in the field of Bayesian nonparametrics for hierarchical modeling (Blei et al., 2004; Teh et al., 2005), clustering (Qin, 2006), natural language processing (Blei et al., 2010), modeling sequential data and network data (Blei and Frazier, 2011), etc.

1.2 Polya Trees

1.2.1 Introduction to the Polya Tree

Dirichlet process and mixtures of Dirichlet processes have yielded much research, extension and application over the last several years. The Dirichlet process, as noted above, has several desirable qualities, particularly for clustering, but the Dirichlet process is constructed to give discrete distributions with probability one; this is prob-

lematic in cases where the data is best modeled in a continuous setting, noting that in the discrete case we expect to see observations repeat.

A prior class that includes the Dirichlet process as a special case is the Polya tree prior (Lavine, 1992); Polya trees generalize the Dirichlet process by allowing the positive mass on the set of continuous distributions. The Polya tree prior was initially summarized by Ferguson (1974) who discusses its construction as a prior distribution for use in deriving Bayesian decision rules in the nonparametric setting. Specifically, Ferguson (1974) explains the dyadic tree structure that splits the support on the real line by subsequent binary splits as depicted in Figure 1.3 for data uniformly distributed on $(0, 1]$ where $Y_{j,k}$ is the probability of seeing an observation on partition k at level j. In the next chapter we define this partitioning scheme for Gaussian distributed data in the context of multiple testing.

All $Y_{j,\cdot}$ are required to be independent between rows (Ferguson, 1974) and $Y_{j,2k} = 1 - Y_{j,2k-1}$ for k, the set of odd numbers from 1 to $2^j - 1$. Each $Y_{j,2k-1}$ is drawn independently from beta distributions whose parameters depend on the number of observations that fall in each piece of subsequent partitions at level $j+1$, i.e. $Y_{j,2k-1} \sim beta(\alpha_{j,k}, \beta_{j,k})$; more details on choosing α and β are provided in the different applications that follow. The probability of being on any interval a is then the product of the $Y_{j,k}$ passed on the path to interval a; i.e. continuing with the example in Figure 1.3: $p\{(7/8, 1]\} = Y_{1,2}Y_{2,4}Y_{3,8}$. It should be noted that the partition sets decrease to $\emptyset$ as J approaches infinity and this allows the probability, defined as a product, to converge to zero with probability one when α and β are appropriately selected. Mauldin et al. (1992) provide conditions where the Polya tree is continuous or absolutely continuous with probability one.

Polya trees have desirable attributes for a prior in that they are conjugate and easily updated (Ferguson, 1974). Specifically, Lavine (1992) explains that a Polya tree prior can be drawn out by questions about $Y_{j,k}$ and shows that the predictive

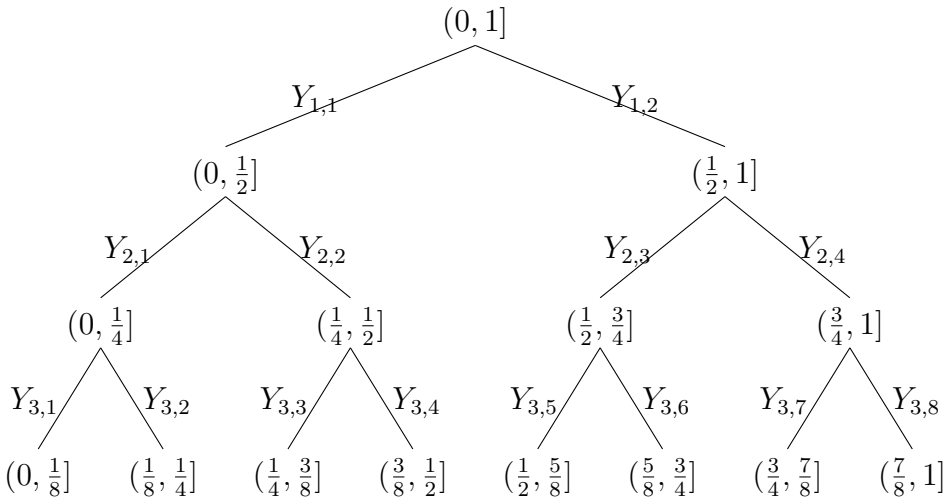

Figure 1.3: Depiction of tree structure from Ferguson (1974).

and posterior distributions can be set up such that one can choose the predictive distribution of the Polya tree. An issue that arises with the Polya tree prior is that the posteriors suffer from discontinuities at the partition end points.

Let G be a Polya tree, i.e. $G \sim PT(c, J, G_0)$ with base distribution G_0 and positive parameter c which controls how "close" the Polya tree is to the baseline distribution. Comparing to the result from the Dirichlet process in Figure 1.1, consider G_0 to be Gaussian, i.e. $G \sim PT(c, J, N(0, 10^2))$. A draw from a Polya tree with $J = 10$ and $c = 1$ can be seen in Figure 1.4; compared to the draw from the Dirichlet process it is clear that the Polya tree provided a larger variety of sample values.

A solution to the issue of discontinuities at the partition end points is to extend the Polya tree to a mixture of Polya trees by considering a Polya tree with a random centering distribution. With mixtures of Polya trees the dependence on the partitions is mitigated by smoothing out the partition discontinuities and the updated Polya tree affords continuous predictive distributions. This idea is not dissimilar to extension of the Dirichlet process to the Dirichlet process mixture model above. Lavine (1994) explores using the Polya tree and mixtures of Polya trees in place of Dirichlet processes and Dirichlet process mixture models in applications including statistical modeling,

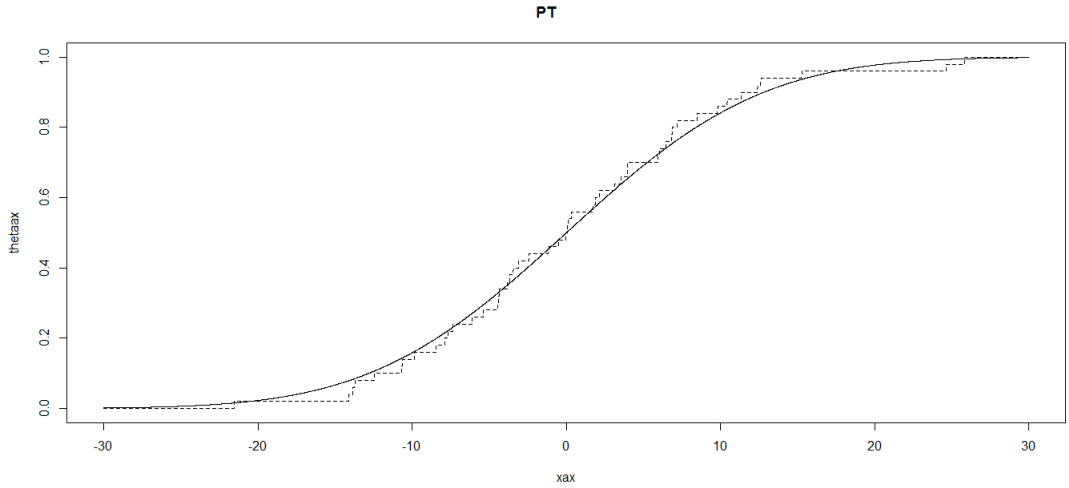

Figure 1.4: Polya tree sample from $N(0, 10^2)$ along with the Gaussian CDF.

modeling errors in regression problems and empirical Bayes problems with much success. A mixture of Polya trees model does quite well in fitting the same $G_0 = N(0, 10^2)$ as seen in Figure 1.5 using just a random sample of $n = 50$ points drawn from G_0.

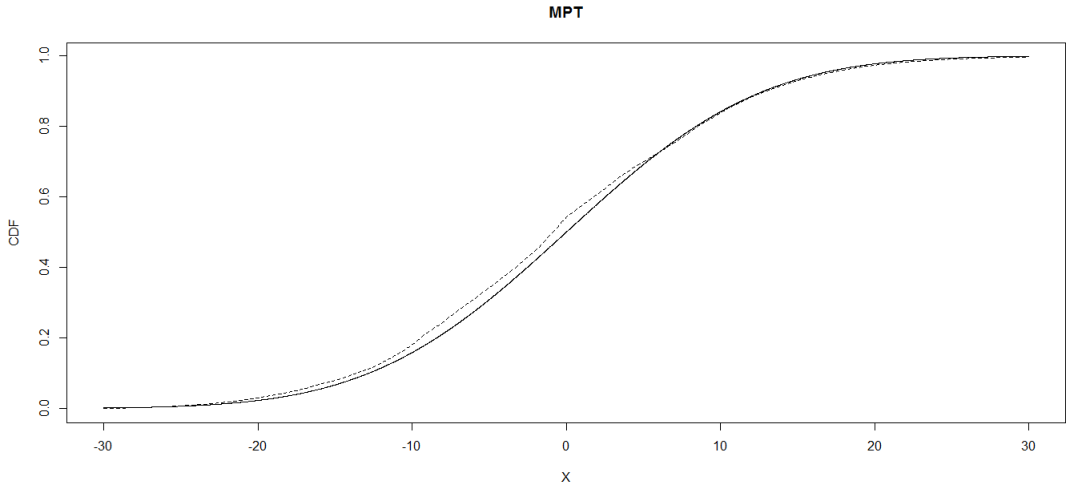

Figure 1.5: Polya tree draw from the $N(0, 10^2)$ cumulative distribution function.

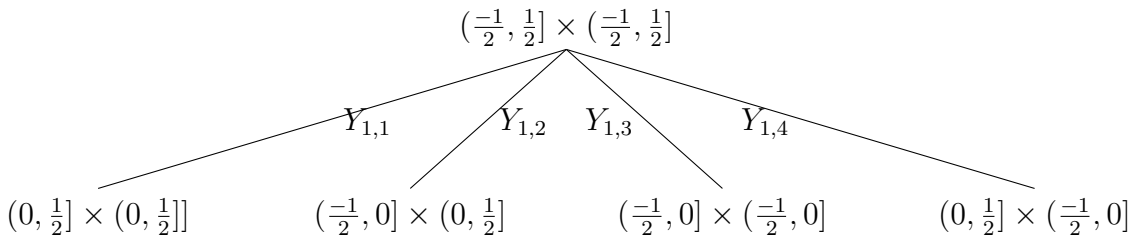

Figure 1.6: Depiction of tree structure from in two dimensions.

1.2.2 Multivariate Polya Trees

Polya trees are easily extended to the multivariate case by simply defining a similar but d-dimensional partitioning structure, depicted in Figure 1.6 for data uniformly distributed on the unit square $(\frac{-1}{2}, \frac{1}{2}] \times (\frac{-1}{2}, \frac{1}{2}]$. In the third chapter we define this partitioning scheme for d-dimensional Gaussian distributed data in the context of supervised learning classification. Note that the beta distributed probabilities Y become Dirichlet and that Figure 1.6 shows that keeping track of the partitioning sequences becomes increasingly difficult as dimensionality increases. Jara et al. (2009) propose a convenient indexing that is demonstrated in Chapter 3.

Paddock et al. (2003) introduced a randomized multivariate Poyla tree which is smoothed over partitions using a random "jitter;" Hanson (2006), rather, introduced a location-scale mixture of Polya trees which directly generalizes the univariate mixture of Polya trees; Hanson et al. (2008) used multivariate Polya trees to model receiver operating characteristic (ROC) curves for evaluating diagnostic test accuracy; Jara et al. (2009) proposed using multivariate Polya trees in generalized linear mixed effect models to remedy the case when the assumption that random effects terms follow a multivariate Gaussian distribution is faulty; Hanson et al. (2011) developed a simple, computationally cheap sampling method for exploring multivariate densities. For more, Müller and Rodriguez (2013) provide a nice summary of various versions of Polya trees.

Chapter 2

Bayesian Nonparametric Multiple Testing

Multiple testing, or multiplicity problems often require testing several means with the assumption of rejecting infrequently, as motivated by the need to analyze DNA microarray data. The goal is to keep the combined rate of false discoveries and non-discoveries as small as possible. A discrete approximation to a Polya tree prior that enjoys fast, conjugate updating, centered at the usual Gaussian distribution is proposed. This new technique and the advantages of this approach are demonstrated using extensive simulation and data analysis accompanied by a Java web application. The numerical studies demonstrate that this new procedure shows promising false discovery rate and estimation of key values in the mixture model with very reasonable computational speed.

2.1 Introduction

In the DNA microarray setting, multiple testing problems often require testing multiple hypotheses H_0: $\theta_i = 0$. The θ_i measurements can contain a range of values, both positive and negative, and the goal is to detect the $|\theta_i|$ that are large enough to reject H_0: $\theta_i = 0$. Initially we consider y_i, $i = 1, \ldots, n$ such that each y_i is independently Gaussian distributed with mean θ_i and fixed, unknown variance σ^2, as considered by Scott and Berger (2006); we follow with a model that accounts for known variances σ_i^2, as considered by Sun and McLain (2012).

Two tasks are considered while testing H_0: $\theta_i = 0$: the task of finding θ_i that are different from zero, and that are "interestingly different" from zero. In the DNA mi-

croarray setting, often y_i are difference measurements of gene expression in two states where low differences are of little interest; scientists often consider genes that show at least two-fold change in expression levels to be differentially expressed, meaning observations with large deviations from the null are sought. It is important to distinguish that the proposed model facilitates hypothesis tests that θ_i are different from zero, but we provide adapted methodology to answer the question of "interestingly different" observations.

2.1.1 Biological Motivation

A living organism's basic make up is due to its genome for all plants, animals and humans. Genomes are comprised of one or many strands of DNA. Each living species, and the individual variations within those species, are defined through the details of DNA. A single cell of a living organism contains at least one copy DNA which is organized into chromosomes; humans have twenty three pairs of chromosomes, most famously the paired X and Y sex chromosomes. The chromosomes, which make up DNA, contain regions called genes that are involved in the production of proteins; each chromosome can contain a different collection of genes.

Genes are important to medical research due to their role in the production of proteins which make up the entirety of organisms by controlling replication, form and mutation. When you look at a human, for instance, everything you see consists of protein: hair, skin, eyes, etc. The rate of cell production is also controlled by genes so the regeneration of skin cells and hair growth etc. are also attributed to genes. Sometimes, though rarely, a genetic mutation can cause a disease like Progeria which causes accelerated aging as illustrated in the 1996 film Jack starring the late Robin Williams.

The effect that genes have on protein production, and consequently an organism's properties, make their influence important. It is often desired to know the difference of

gene expression, or activity, in different stages; as the levels of gene expression vary, the production of protein they manufacture may change possibly yielding physical abnormalities. Perhaps the most infamous example is when cells grow and divide at an abnormal pace creating growths within the body; many times these growths are benign, a non-cancerous growth, and other times they are metastasized, a cancerous growth. In much of cancer research, expression values of genes that regulate cell growth are important - both in susceptibility determination (van 't Veer et al., 2002) and the development of personalized gene therapy as well as the classification of types of cancer (Golub et al., 1999). It should be noted that many other medical issues are explored using similar gene expression research such as Autism (Liu et al., 2014), Gulf War illness (Craddock et al., 2015), type 2 diabetes (Patti et al., 2003), etc.

DNA microarray data displays sets of microscopic spots of DNA laid out on a piece of glass; together this is referred to as a DNA chip. DNA chips come in two flavors, cDNA or oligonucleotides. cDNA requires that the DNA chip have full-length transcripts printed onto the glass and oligonucleotides are chemically synthesized on the glass and then exposed to probes which extract information from a particular cell during different stages. Scientists usually prefer oligonucleotides because these nucleotides can identify which genes are active at certain stages of development, or under certain environmental conditions. The goal is often to decide which genes, and subsequently which proteins, are active or inactive during these stages or environments. Before DNA microarray data was available, traditional methodologies only provided ways to study one or a few genes at a time; now widespread availability makes it possible to explore all the genes in a single experiment.

Efron et al. (2001) and Do et al. (2005) considered a prevalent example of multiple testing when analyzing DNA microarray data. The problem explored was how certain treatments, environment changes or disease affect gene expression. The DNA microarray data of Efron et al. (2001) consists of 6,810 genes exposed to eight condi-

tions. Even if all the genes were not activated, meaning each test H_0: $\theta_i = 0$ should result in a failure to reject, setting up the hypothesis tests with a significance level 0.05 leads to making a type I error roughly five percent of the time; if this were the case when testing 6,810 genes one would expect to make type I error about 340 times.

2.1.2 Motivations Beyond Biology

Research in multiple testing has been heavily motivated by DNA microarray data. The need for developing methods that accurately control for multiple testing should increase as the world more heavily depends on big data solutions in both science and business, which are becoming more prevalent. Multiple testing can be useful in multiple comparison problems including brain imaging (Chumbley and Friston, 2009), financial analysis (Bajgrowicz and Scaillet, 2007), market research (Blomquist, 2014) and many other data heavy disciplines that exist today. Sun and McLain (2012) provide another prominent example affected by heteroscedasticity; educational survey data from the Adequate Yearly Progress (AYP) study on the academic performances of students across different social and fiscal demographics. These data are reanalyzed in Section 7.

2.1.3 Paper Outline

The proposed models for multiple testing are outlined in Section 2 and outlines for the Gibbs sampling procedure for the common and known variance cases in Section 3 and Section 4 respectively. In Section 5 we explore proposed test statistics for the multiple testing procedure as well as details about the false discovery rate. The success and advantages of this new approach are demonstrated through extensive simulation in Section 6 and real data analyses in Section 7.

2.2 Models

Consider median-zero data $\mathbf{y} = (y_1, \ldots, y_n)'$ with mean vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$ modeled

$$y_i|\boldsymbol{\theta}, \sigma^2 \overset{ind.}{\sim} N(\theta_i, \sigma^2), \quad \theta_i|w, G \overset{ind.}{\sim} wg(\theta) + (1-w)\delta_0(\theta), \tag{2.1}$$

where $G(\cdot)$ is a cumulative distribution function with corresponding density $g(\cdot)$, w is the mass parameter in the mixture model denoting the proportion of non-null observations, and δ_x is the Dirac measure at x. Scott and Berger (2006) consider a parametric case where G is Gaussian with mean zero and standard deviation v, i.e. $G = N(0, v^2)$.

The resulting marginal model can be written

$$\begin{aligned} y_i &\sim w\int_{-\infty}^{\infty} N(\theta, \sigma^2)G_1(d\theta) + (1-w)\int_{-\infty}^{\infty} N(\theta, \sigma^2)G_0(d\theta) \\ &= w\int_{-\infty}^{\infty} N(\theta, \sigma^2)G_1(d\theta) + (1-w)N(0, \sigma^2), \end{aligned}$$

where $G_0 = \delta_0$. Note that we generalize only the distribution of the non-null means assumed in Scott and Berger (2006). Do et al. (2005) consider a further generalization of this model for differentially expressed genes where G_1 and G_0 have Dirichlet process priors, namely $G_1 \sim DP(\alpha_1, G_1^*)$, $G_1^* = 0.5N(-b_1\sigma_1^2) + 0.5(b_1, \sigma_1^2)$, and $G_0 \sim DP(\alpha_0, G_0^*)$ where $G_0^* = N(0, \sigma_0^2)$.

Bogdan et al. (2008) instead consider a nonparametric generalization of the marginal model on the $y_1, \ldots, y_n$ proposed by Scott and Berger (2006), namely

$$y_i \sim \int_{-\infty}^{\infty} N(\theta, \sigma^2)G(d\theta), \tag{2.2}$$

where $G \sim DP(\alpha, wN(0, v^2) + (1-w)\delta_0)$. This would appear to be the same model, but note then that $G(\{0\})|w \sim \text{beta}(\alpha(1-w), \alpha w)$. Assuming $w \sim \text{beta}(a_w, b_w)$ independent of $\alpha \sim \exp(a_\alpha)$ (Bogdan et al. take $a_w = 1$, $b_w = 22.76$, and $a_\alpha = 1$) the prior probability q of a null has density

$$P(\theta = 0) \sim \pi(q) = \int_0^{\infty} a_\alpha \exp(-a_\alpha\alpha)\int_0^1 \beta(q|\alpha(1-w), \alpha w)\beta(w|a_w, b_w)dwd\alpha.$$

The prior probability of a null is no longer beta(a_w, b_w); placing a nonparametric prior directly on the non-null distribution, as in (2.1), retains the interpretation of w as the probability of a non-null. Guindani et al. (2009) also considered essentially

the same model as (2.2), but from a decision-theoretic framework. Dahl et al. (2009) generalize (2.2) for multivariate nonparametric random effects models. Martin and Tokdar (2012) modify (2.2) to guarantee identifiability,

$$y_i \sim \int_{-\infty}^{\infty} N(\mu + \tau\sigma\theta, \sigma^2) G(d\theta),$$

where $\tau > 1$ is assigned a shifted standard log-normal prior and G is instead estimated using predictive recursion to avoid difficulties stemming from the usual Dirichlet process prior. Dirichlet process distributions are almost surely discrete and computation does not scale well with the number of observations, i.e. large datasets with several hundreds of thousands of observations are infeasible.

A prior class that includes the Dirichlet process as a special case is the Polya tree prior (Lavine, 1992). This paper proposes to model the non-null distribution G in (2.1) with a Polya tree prior, keeping the dyadic tree structure of Ferguson (1974) and Lavine (1992) on the conditional probabilities, but to terminate the tree at a finite level J and simply replace the sets at this level with point mass. The idea is simple but results in a nonparametric distribution that is exceptionally computationally tractable with many of the benefits of using a Dirichlet process while maintaining simple interpretation of the parameters and posterior distributions.

We generalize the model of Scott and Berger (2006) and offer an alternative nonparametric approach to Martin and Tokdar (2012) by assigning G a nonparametric finite Polya tree prior with J levels that is centered at a $N(0, v^2)$ distribution. That is,

$$G|c, v^2 \sim PT_J\{c, N(0, v^2)\}, \tag{2.3}$$

where PT_J denotes a finite Polya tree with J levels, and c is a parameter that controls how "close" G is to the centering distribution $N(0, v^2)$. The variance components σ^2 and v^2 are given the improper prior $p(\sigma^2, v^2) \propto (v^2 + \sigma^2)^{-2}$ of Scott and Berger (2006); note that this results in a conditionally proper prior for $[v^2|\sigma^2]$, which is needed as there may be no non-zero θ_i with positive probability $(1 - w)^n$. The prior on w, the amount of mass assigned to $g(\theta)$, is beta, i.e. $w \sim \text{beta}(a_w, b_w)$. For the special case $G = N(0, v^2)$ Scott and Berger (2006) set $a_w = \alpha + 1$ and

$b_w = 1$. Setting $\alpha = 0$ yields a uniform prior in the case where no prior information is available. In the case where prior information about the non-null proportion $\hat{p}$ is provided, setting $\alpha = \log 0.5 / \log \hat{p}$ makes $\hat{p}$ the prior median with reasonable variation. In some experiments experts can provide more detailed information on w and its variation. Consider taking $a_w = \hat{p}\, m$ and $b_w = (1 - \hat{p})\, m$, where $\hat{p} \in (0, 1)$ and $m > 0$ are constants such that $E(w) = \hat{p}$ and $var(w) = \hat{p}\,(1 - \hat{p}) \,/\, (m + 1)$, matching the information provided by the expert.

Deciphering which observations are "interestingly different" from zero has much to do with the prior on w, the amount of mass assigned to $g(\theta)$ or the proportion of non-null observations. By altering the prior on w we allow the user to indicate how many of the observations should be rejected, allowing the tweaking of w to be interpreted as the proportion of "interestingly different" observations instead of just observations with non-zero means.

By default, this model completes a multiple testing procedure for point mass at zero. Many times "interestingly different" observations are desired, particularly in microarray analysis where investigating genes via genetic testing can be an expensive endeavor. Consider the case where $w = w_0$ is the result with the default uniform prior. When observations that are "interestingly different" than zero are desired, one can then try using prior information $\hat{p} < w_0$ and m so that $E(w) = \hat{p}$ and $var(w)$ is small, e.g. $var(w) \leq .01$. Less, but "more interesting" rejections can be found by systematically decreasing $\hat{p}$. This is demonstrated in the simulations and data analyses of Section 6 and Section 7.

An alternative methodology for finding "interestingly different" observations is provided in Section 5, where test statistics and thresholds are introduced, by implementing a relative cost, κ, which intimates whether a false discovery or false non-discovery is more costly. When it is more expensive to make a false discovery the model completes Gibbs sampling with the same flat prior on w, unless otherwise

stated, but the posterior analysis would reveal similar results where less, but "more interesting" rejections are revealed.

The Polya tree prior was initially summarized by Ferguson (1974), and further developed by Lavine (1992, 1994), and Mauldin et al. (1992). Hanson (2006) discusses inference for mixtures of finite Polya trees, which smooth out the effect of the partition on posterior inference. Briefly, the prior (2.3) on G adds to $N(0, v^2)$. The $2^J - 1$ conditional probabilities that refine G's shape are

$$Y_{j,k}|c \overset{ind.}{\sim} \text{beta}(cj^2, cj^2),$$

where $j = 1, \ldots, J$, and k are the odd numbers from 1 to $2^j - 1$ at any level j. For any $Y_{j,k}$ where k is odd, let $Y_{j,k+1} = 1 - Y_{j,k}$. Define $\mathcal{Y} = \{Y_{jk} : j = 1, \ldots, J; k = 1, \ldots, 2^j\}$. A Polya tree parameter is the conditional probability $Y_{j,k} = G\{B_v(j,k)|B_v(j-1, \lceil k/2 \rceil)\}$ where $B_v(j,k) = (v\Phi^{-1}((k-1)/2^j), v\Phi^{-1}(k/2^j))$, is the interval for the partition k on level j. Note that $B_v(j,1), \ldots, B_v(j,2^j)$ partitions $\mathbb{R}$ up to a set of measure zero for each $j = 1, \ldots, J$, and for any measurable $A \subset \mathbb{R}$, $E\{G(A)\} = \int_A \phi(t|0, v^2)dt$ where $\phi(\cdot|\mu, \sigma^2)$ is the density of a Gaussian random variable with mean μ and standard deviation σ. The probability of being in set k at level J is

$$p_{\mathcal{Y}}(k) = \prod_{j=1}^{J} Y_{j,\lceil k2^{j-J} \rceil}, \quad k = 1, \ldots, 2^J,$$

with $\lceil \cdot \rceil$ the usual ceiling function. For example, for $J = 5$ one can find $p_{\mathcal{Y}}(11)$ with

$$\begin{aligned} p_{\mathcal{Y}}(k = 11) &= \prod_{j=1}^{J=5} Y_{j,\lceil (11)2^{j-J} \rceil} \\ &= Y_{1,\lceil (11)2^{1-5} \rceil} Y_{2,\lceil (11)2^{2-5} \rceil} Y_{3,\lceil (11)2^{3-5} \rceil} Y_{4,\lceil (11)2^{4-5} \rceil} Y_{5,\lceil 112^{5-5} \rceil} \\ &= Y_{1,1} Y_{2,2} Y_{3,3} Y_{4,6} Y_{5,11}; \end{aligned} \tag{2.4}$$

see the highlights in Figure 2.1.

Although the measure G is discrete, the usual density estimate from a mixture of Polya trees can be used to smooth G and give an idea of how mass is spread out. Another option is to simply plot the location and height of the point masses. We opt for the former using results from Hanson (2006)(see formula (6)). For $G \sim PT_J(c, N(0, v^2))$, the density $g(\theta) = g(\theta|\mathcal{Y}, v^2)$ of $G(\theta)$ given $\mathcal{Y}$ and v has the following form:

$$g(\theta|\mathcal{Y}, v^2) = 2^J p\{k_v(\theta)\}\phi(\theta|0, v^2) = 2^J \phi(\theta|0, v^2) \sum_{k=1}^{2^J} I\{k_v(\theta) = k\} p_{\mathcal{Y}}(k),$$

where 2^J gives the number of partitions in the last level of the Polya tree and $k_v(\theta) = \lceil 2^J \Phi(\theta/v) \rceil$, gives k such that $\theta \in B_v(J,k)$.

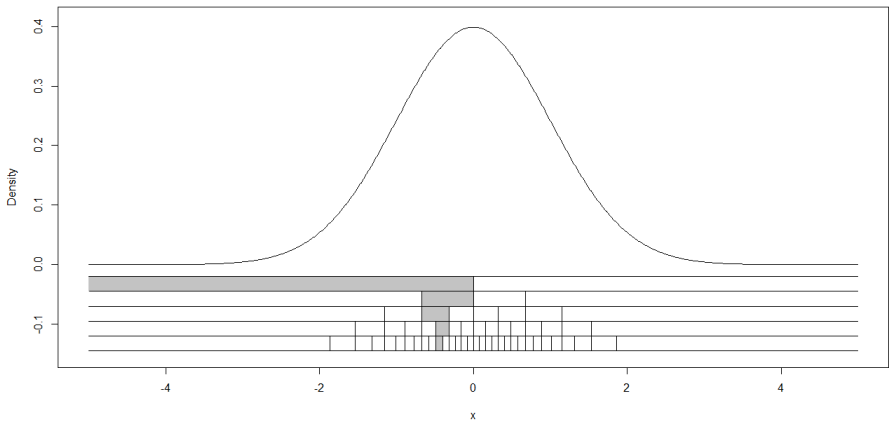

Figure 2.1: Graph of Gaussian pdf with Polya tree partitions and example highlights for $p_{\mathcal{Y}}(k = 11)$.

2.2.1 Direct inference via adaptive MCMC

Following Scott and Berger (2006), first consider marginalized inference. Marginalizing over θ, each datum arises independently from the density

$$f(y_i|\sigma^2, w, G) = w \int_{\mathbb{R}} g(\theta)\phi(y_i|\theta, \sigma^2)d\theta + (1-w)\phi(y_i|0, \sigma^2).$$

For $a < b$ note the identity

$$\int_a^b \phi(y|\theta, \sigma^2)\phi(\theta|0, v^2)d\theta = \phi(y|0, \sigma^2 + v^2)\left[\Phi\left\{\frac{b(\sigma^2+v^2) - v^2 y}{\frac{\sigma v}{\sqrt{\sigma^2+v^2}}}\right\} - \Phi\left\{\frac{a(\sigma^2+v^2) - v^2 y}{\frac{\sigma v}{\sqrt{\sigma^2+v^2}}}\right\}\right].$$

This leads to the density of y given v, σ, and $\mathcal{Y}$ as

$$\begin{aligned}
m(y|v, \sigma, \mathcal{Y}, w) &= w \int_{\mathbb{R}} g(\theta|\mathcal{Y}, v)\phi(y|\theta, \sigma^2)d\theta + (1-w)\phi(y_i|0, \sigma^2) \\
&= w \int_{\mathbb{R}} 2^J \phi(\theta|0, v^2)\left[\sum_{k=1}^{2^J} I\{k_v(\theta) = k\}p_{\mathcal{Y}}(k)\right]\phi(y|\theta, \sigma^2)d\theta + (1-w)\phi(y_i|0, \sigma^2) \\
&= w \int_{\mathbb{R}} 2^J \left[\sum_{k=1}^{2^J} I\{k_v(\theta) = k\}p_{\mathcal{Y}}(k)\right]\left[\phi(y|\theta, \sigma^2)\phi(\theta|0, v^2)\right]d\theta + (1-w)\phi(y_i|0, \sigma^2) \\
&= w2^J\phi(y|0, \sigma^2 + v^2)\sum_{k=1}^{2^J} p_{\mathcal{Y}}(k)\Delta(y, k|\sigma, v) + (1-w)\phi(y_i|0, \sigma^2).
\end{aligned}$$

where

$$\Delta(y, k|v, \sigma) = \Phi\left\{\frac{\Phi^{-1}\left(\frac{k}{2^J}\right)(\sigma^2+v^2) - y}{\frac{\sigma}{\sqrt{\sigma^2+v^2}}}\right\} - \Phi\left\{\frac{\Phi^{-1}\left(\frac{k-1}{2^J}\right)(\sigma^2+v^2) - y}{\frac{\sigma}{\sqrt{\sigma^2+v^2}}}\right\}.$$

The unnormalized posterior density is then

$$\begin{aligned} p(v,\sigma,c,\mathcal{Y},w) &\propto \left[\prod_{i=1}^{n} m(y_i|v,\sigma,\mathcal{Y},w)\right] p(v)p(\sigma)p(c) \\ &\quad \times \left[\prod_{j=1}^{J}\prod_{k=1}^{2^{j-1}} \text{beta}(Y_{j,2k-1}|cj^2,cj^2)\right] \text{beta}(w|a_w,b_w). \end{aligned}$$

The dimension of the posterior parameter vector is $2^J + 3$. Adaptive Metropolis-Hastings (Haario et al., 2001) can be used here to obtain posterior inference.

2.3 Discrete Approximation to Simplify Posterior Updating

2.3.1 Discrete approximation

To simplify the computational complexity, consider a discrete approximation to the Polya tree. Define G to be the finite discrete measure

$$G(\cdot) = \sum_{k=1}^{2^J} p_{\mathcal{Y}}(k)\delta_{\theta_k}(\cdot), \quad \theta_k = v\Phi^{-1}\left(\frac{k-0.5}{2^J}\right) \stackrel{def}{=} vt_k. \tag{2.5}$$

Note that as J gets large, the intervals $B_v(J,k)$ get smaller, except in the tails, and $g(\cdot)$ varies less over the intervals. Here $g(\cdot)$ follows $N(0,v^2)$ over each interval k, and can be approximated with just one "representative" point, the mid-quantile, θ_k, plus the associated probability $p_{\mathcal{Y}}(k)$ of the interval under G. This leads to the density of y_i given $\sigma^2, w, v, \mathcal{Y}$ as

$$\begin{aligned} f(y_i|\sigma^2,w,v,\mathcal{Y}) &= w\int_{\mathbb{R}} \phi(y_i|\theta,\sigma^2)G(d\theta) + (1-w)\phi(y_i|0,\sigma^2) \\ &= w\sum_{k=1}^{2^J} p_{\mathcal{Y}}(k)\phi(y_i|\theta_k,\sigma^2) + (1-w)\phi(y_i|0,\sigma^2). \end{aligned}$$

Note that one can marginalize over σ^2 with

$$\int_0^\infty \phi(y_i|\theta,\sigma^2)\Gamma(\sigma^{-2}|a,b)d\sigma^{-2} = \frac{b^a}{\sqrt{2}\beta(a,1/2)}\left[b + \left(\frac{y_i-\theta}{\sqrt{2}}\right)^2\right]^{-1/2-a}.$$

This can form the basis of inference using a "black box" sampler, for example an adaptive Metropolis-Hastings proposal (Haario et al., 2001).

2.3.2 Gibbs sampling through data augmentation

To simplify computation, component membership indicators are introduced. Let $z_i = j$ iff $y_i \sim N(\theta_j,\sigma^2)$, where $\theta_0 = 0$ and $j = 0,1,\ldots,2^J$. Then

$$P(z_i = k|v,\sigma^2,w,\mathcal{Y}) \propto \left\{ \begin{array}{ll} \phi(y_i|0,\sigma^2)(1-w) & k=0 \\ \phi(y_i|t_kv,\sigma^2)wp_{\mathcal{Y}}(k) & k>0 \end{array} \right\}.$$

$$f(v|\sigma^2, \mathbf{z}) \propto f(v|\sigma^2) \prod_{i:z_i \neq 0} \exp\{-0.5\sigma^{-2}(y_i - t_{z_i} v)^2\},$$

so then

$$f(v|\sigma^2, \mathbf{z}) \propto N\left(v| \frac{\sum_{i=1}^n t_{z_i} y_i}{\sum_{i=1}^n t_{z_i}^2}, \frac{\sigma^2}{\sum_{i=1}^n t_{z_i}^2}\right) f(v|\sigma^2) = N\left(v| \frac{\sum_{i:z_i>0} t_{z_i} y_i}{\sum_{i:z_i>0} t_{z_i}^2}, \frac{\sigma^2}{\sum_{i:z_i>0} t_{z_i}^2}\right) f(v|\sigma^2).$$

Here, v^* is sampled from the Gaussian full conditional (under a flat prior) above and accepted with Metropolis-Hastings probability $1 \wedge f(v^*|\sigma^2)/f(v|\sigma^2)$, where $f(v|\sigma^2)$ is the induced prior from Scott and Berger (2006). Note that the conditional distribution depends only on those observations for which $z_i > 0$, i.e. those observations deemed to be from the non-null distribution. During the Gibbs sampler one can sample and keep track of the z_i while simultaneously updating v and the means $t_k v$. Also note the non-identifiability of the model: different values of $\mathcal{Y}$ and v can give the same likelihood: $v \to -v$ and $p_{\mathcal{Y}}(k) \to p_{\mathcal{Y}}(2^J - j + 1)$. Thus $v > 0$ is needed and maintains interpretation of v as a scale parameter.

The full conditional distributions for w and σ^{-2} are

$$w|\mathbf{z} \sim \text{beta}\left(a_w + \sum_{i=1}^n I\{z_i > 0\}, b_w + \sum_{i=1}^n I\{z_i = 0\}\right),$$

with $a_w = \alpha + 1$ and $b_w = 1$, and

$$\sigma^{-2}|\mathbf{z}, v \sim \Gamma\left(a_\sigma + 0.5n, b_\sigma + 0.5\sum_{i=1}^n (y_i - v t_{z_i})^2\right) f(\sigma^{-2}|v^2),$$

where $t_0 = 0$. Here σ^{-2} is drawn from the gamma proposal above and accepted with a Metropolis-Hastings probability similar to $[v|\mathbf{z}, \sigma^2]$ above. Let

$$n(J, k) = \sum_{i=1}^n I\{z_i = k\}, \quad n(j-1, k) = n(j, 2k-1) + n(j, 2k), j = 2, \ldots, J,$$

then

$$Y_{j,2k-1} \sim \text{beta}(c\rho(j) + n(j, 2k-1), c\rho(j) + n(j, 2k)), \ k = 1, \ldots, 2^{j-1}, \ j = 1, \ldots, J.$$

Note that $Y_{1,1} = 0.5$ and is not sampled; this ensures that zero is a median of G.

Finally,

$$p_{\mathcal{Y}}(k) = \prod_{j=1}^J Y_{j, \lceil k \ 2^{j-J} \rceil}, \ k = 1, \ldots, 2^J.$$

The probabilities $\mathcal{Y}$ are "unhinged" from the location of the sets, or points θ_k, making MCMC sampling easy.

2.4 Deconvolution with Known Variances

Now consider the situation where the y_i are observed with known heteroscedastic error σ_i^2:

$$y_i|\boldsymbol{\theta}, \sigma^2 \overset{ind.}{\sim} N(\theta_i, \sigma_i^2),$$

as considered by Sun and McLain (2012). For this model, σ^2 no longer needs to be updated. Component membership is updated via

$$P(z_i = k|v, \sigma_i^2, w, \mathcal{Y}) \propto \left\{ \begin{array}{ll} \phi(y_i|0, \sigma_i^2)(1-w) & k = 0 \\ \phi(y_i|t_k v, \sigma_i^2) w p_{\mathcal{Y}}(k) & k > 0 \end{array} \right\}.$$

The full conditional for v is now updated to

$$f(v|\boldsymbol{\sigma}, \mathbf{z}) \propto f(v) \prod_{i:z_i \neq 0} \exp\{-0.5\sigma_i^{-2}(y_i - t_{z_i} v)^2\},$$

where

$$\begin{aligned} f(v|\boldsymbol{\sigma}, \mathbf{z}) &\propto N\left(v|\frac{\sum_{i=1}^n t_{z_i} y_i/\sigma_i^2}{\sum_{i=1}^n t_{z_i}^2/\sigma_i^2}, \frac{1}{\sum_{i=1}^n t_{z_i}^2/\sigma_i^2}\right) f(v) \\ &= N\left(v|\frac{\sum_{i:z_i>0} t_{z_i} y_i/\sigma_i^2}{\sum_{i:z_i>0} t_{z_i}^2/\sigma_i^2}, \frac{1}{\sum_{i:z_i>0} t_{z_i}^2/\sigma_i^2}\right) f(v). \end{aligned}$$

In what follows, simply take $f(v) \propto I_{(0,u)}(v)$ for some large $u > 0$.

It should be noted that setting c and J large (e.g. $c = 10,000$ and $J = 8$) provides a good approximation to Scott and Berger (2006). For a fixed J and v, $c \to \infty$ implies that the cumulative distribution function $G(\theta_k) \to \Phi(\theta_k/v)$, i.e. converges to the Gaussian $N(0, v^2)$ centering distribution value. The points $\{\theta_k\}$ become more dense as $J \to \infty$ and, in fact, the partition sets generate the Borel sets $\mathcal{B}(\mathbb{R})$. As a consequence $c \to \infty$ and $J \to \infty$ imply that G converges in distribution to $N(0, v^2)$. In several analyses simply taking $J = 8$ has yielded no difference in the log pseudo marginal likelihood (LPML) (Gelfand and Dey, 1994), a predictive measure of model fit, between a finite Polya tree prior and the discrete approximation proposed here.

A referee asked about an extension to different but unknown variances from a common distribution, i.e. a hierarchical model. Such an extension is straightforward. For example, assume $\sigma_i^{-2}|\lambda \overset{iid}{\sim} \exp(\lambda)$ and $\lambda \sim \exp(\lambda_0)$ where $E(\lambda) = 1/\lambda_0$. Then,

conditionally $\sigma_i^{-2}|\text{else} \sim \Gamma(.5, \lambda + .5(y_i - t_{z_i}v)^2)$ and $\lambda|\text{else} \sim \Gamma(n+1, \lambda_0 + \sum_{i=1}^n \sigma_i^{-2})$; updating is easy. Integrating out $\sigma_1^{-2}, \ldots, \sigma_n^{-2}$ is the same as assuming $[y_i|z_i, v, \lambda] \sim t_{z_i}v + \sqrt{\lambda}T_2$ where T_2 is a student-t random variable with 2 degrees of freedom, i.e. the kernel for y_i changes from Gaussian to student-t with infinite variance. In general, a hierarchical prior on the variances simply changes the kernel. One intriguing possibility is a Dirichlet process prior on the $\sigma_1^{-2}, \ldots, \sigma_n^{-2}$, which would force clusters of observations to have the same variance, but different variances across clusters.

2.5 Error Rates

With single hypothesis tests the rejection threshold is moved to control type I error. In the case of multiple testing there may be hundreds, thousands or possibly millions of hypothesis tests each having their own type I and type II errors. These errors need to be combined to talk about the overall error, i.e. specificity or sensitivity of multiple hypothesis tests. Consider, family wise error rate (FWER) methods, i.e. Sidak (1967), Tukey (1949), and Hochberg (1988), the probability of making one or more type I errors. The goal is to limit this probability at most to a fixed value.

$$\begin{aligned} FWER &= P(\text{False Discoveries} \geq 1) \\ &= 1 - P(\text{False Discoveries} = 0) \leq \alpha. \end{aligned}$$

The problem here, in line with one single hypothesis test, is that limiting this probability leads to an increase in type II error rate and a decrease in overall power.

The false discovery rate, FDR, introduced by Benjamini and Hochberg (1995) has become a common error measurement in multiple testing problems. It provides a way to control type I error like FWER, but does not suffer from the low power that occurs with FWER methods. It should also seem more logical to consider the rate or proportion of errors instead of whether or not any errors were made; five false discoveries in ten hypothesis tests is much worse than the same number out of one hundred tests. Müller et al. (2007) suggested a variation of FDR from a conditional Bayesian perspective, but we continue with the traditional measurement.

Many FDR methods start with a calculation of the false discovery proportion, FDP, the proportion of false discoveries among all discoveries, closely aligned with type I error. FDR is then the expected value of FDP. Similarly, the false non-discovery rate, FNR, starts with a calculation of the false non-discovery proportion, FNP, as the proportion of false non-discoveries among all non-discoveries, aligned with type II error. FNR is the expected value of FNP.

The tests of interested are $H_i : \theta_i = 0;\ i = 1, \ldots, n$. Let $\gamma_i = (\gamma_1, \ldots, \gamma_n)$ with $\gamma_i = I(\theta_i \neq 0)$. Let $\tau = (\tau_1, \ldots, \tau_n)$ be the collection of test statistics $\tau_i = P(z_i \neq 0|\mathcal{Y})$, the posterior probability that we fail to reject H_i, with test threshold values of $T = (T_1, \ldots, T_n)$, with default

$$T_{i1} = \frac{\kappa \int_0^\infty |\frac{\theta_i}{v}| \pi(\theta_i|\gamma_i = 0, \boldsymbol{y})\, d\theta_i}{1 + \kappa \int_0^\infty |\frac{\theta_i}{v}| \pi(\theta_i|\gamma_i = 0, \boldsymbol{y})\, d\theta_i} = \frac{\kappa E(|\frac{\theta_i}{v}|)}{1 + \kappa E(|\frac{\theta_i}{v}|)} = \frac{\kappa E(|t_k|)}{1 + \kappa E(|t_k|)},$$

which is based on the threshold proposed by Scott and Berger (2006). The relative cost of making a false non-discovery compared to a false discovery is denoted by κ. All else equal, T_{i1} is increasing in κ indicating that if the cost of a false non-discovery is higher relative to a false discovery, larger κ is desirable.

In many cases, large observations cause this threshold to approach one, regardless of the observations' variances. This causes observations with large $E(|\theta_i|)$ to be labeled as non-discoveries overall even if they were labeled as non-discoveries just once during Gibbs sampling. This is an important consideration because real world observations can be quite large; the carcinoma data from Notterman et al. (2001), considered in the data analysis in Section 7, has difference values over $1,300$ for example. To account for these cases scaling θ_i by the variance term v is required.

In the case where large observations are expected, one can alternatively use the originally proposed threshold of Scott and Berger (2006)

$$T_{i2} = \frac{\kappa \int_0^\infty |\theta_i| \pi(\theta_i|\gamma_i = 0, \boldsymbol{y})\, d\theta_i}{1 + \kappa \int_0^\infty |\theta_i| \pi(\theta_i|\gamma_i = 0, \boldsymbol{y})\, d\theta_i} = \frac{\kappa E(|\theta_i|)}{1 + \kappa E(|\theta_i|)} = \frac{\kappa E(|t_k v|)}{1 + \kappa E(|t_k v|)}.$$

The motivation for the unscaled threshold, T_{i2}, was to create a threshold such that observations with posterior means close to zero have lower thresholds than those with posterior means far from zero. In many cases, small observations cause this threshold

to approach zero, regardless of the observations' variances. This causes observations with small $E(|\theta_i|)$ to be labeled as discoveries overall even if they were labeled as discoveries just once during Gibbs sampling. This deficiency is important to consider because many real world observations can be quite small; it should be noted that this threshold fails in the case of testing the difference of proportions as considered in the data analysis of success rates on math exams of Section 7.

The two thresholds are very comparable when considering observations of magnitude 10^1, T_{i1} performs better in the case when considering the observations of magnitude 10^{-1} or smaller and T_{i2} performs better in the case when considering the observations of magnitude 10^2 or larger. Finally, this gives the decision functions $\delta_{i1} = I(\tau_i > T_{i1})$ and $\delta_{i2} = I(\tau_i > T_{i2})$ which are toggled between depending on the magnitude of observations.

Table 2.1: Errors calculation for multiple hypothesis tests

	H_0 True	H_0 False	Total
Reject H_0	$V = \sum_i (1-\gamma_i)\,\delta_i$	$S = \sum_i \gamma_i \delta_i$	$R = \sum_i \delta_i$
Fail to Reject H_0	$U = \sum_i (1-\delta_i)\,(1-\gamma_i)$	$W = \sum_i (1-\delta_i)\,\gamma_i$	$A = \sum_i (1-\delta_i)$
Total	$\sum_i (1-\gamma_i)$	$\sum_i (\gamma_i)$	n

Table 2.1 summarizes type I and type II errors across multiple hypothesis tests. The number of discoveries can be written as $R = \sum_i \delta_i$, with the number of false discoveries $V = \sum_i (1-\gamma_i)\,\delta_i$ making $FDP = \frac{V}{max(R,1)}$. FDR is then $E[\frac{V}{max(R,1)}]$. The number of non-discoveries can be written as $A = \sum_i (1-\delta_i)$, with the number of false non-discoveries $W = \sum_i (1-\delta_i)\,\gamma_i$ making $FNP = \frac{W}{max(A,1)}$. FNR is then $E[\frac{W}{max(A,1)}]$. In the same vein, the missed discovery rate, MDR, is $E[\frac{W}{max(m_0,1)}]$ where $m_0 = \sum_i \gamma_i$, the number of θ_i that should be rejected. The calculations should clarify the difference of FNR and MDR. FNR is the ratio of false non-discoveries among all non-discoveries whereas MDR is the ratio of false non-discoveries among all θ_i that should be rejected.

FDR and FNR are the expected value of FDP and FNP respectively and a simple Law of Large Numbers argument allows us to approximate FDR and FNR with

FDP and FNP respectively in the simulation studies in Section 6. The marginal versions are defined as $MFDR = \frac{E(V)}{E[max(R,1)]}$ and $MMDR = \frac{E(W)}{E[max(m_0,1)]}$. Genovese and Wasserman (2002) show that MFDR and FDR are negligibly different in large problems and are argued better than their non-marginal counterparts by Storey (2002) and Wu and Cai (2007). A more advanced procedure for error measurement is discussed by Peña et al. (2011) who consider taking the individual powers of each test, which are left out of other methods, into account for both FWER and FDR methods. Their research focuses on Neyman-Pearson Most Powerful tests of simple hypotheses with promising results that indicate individual powers are important to multiple tests. This paper does not include a model that accounts for individual power, but it is an interesting avenue to consider in further research.

2.6 Simulations

Three scenarios are considered for simulation, one where the means are distributed $G_1 = N(0, 2^2)$, another where the means follow a skewed, bimodal, median-zero mixture of two Gaussians $G_2 = 0.4N(-2.62, 1^2)+0.6N(0.48, 0.5^2)$, and another where the means follow a symmetric, bimodal, median-zero mixture of two Gaussians $G_3 = 0.5N(-3, 0.5^2)+0.5N(3, 0.5^2)$. For each G_j, $j = 1, 2, 3$, we simulate $\theta_1, \ldots, \theta_{500} \overset{iid}{\sim} G_j$ and $\theta_i = 0$ for $i = 501, \ldots, 3000$ i.e. $w = 1/6$. Finally, for each G_j we simulate (i) $y_i \overset{ind.}{\sim} N(\theta_i, 0.5^2)$ for the model with a common, unknown σ^2, or (ii) $y_i \overset{ind.}{\sim} N(\theta_i, \sigma_i^2)$ where $\sigma_i \sim \Gamma(5, 10)$ where $E(\sigma_i) = 0.5$. An equal-tailed 95% probability interval for σ_i is $(0.16, 1.02)$. For G_1, G_2 and G_3 one hundred datasets were generated.

In each of the three scenarios we explore the FDR and FNR, across varying values of κ, J, w and prior information on w as well as the estimates of the non-null densities. Then, we increase the sample size to $n = 30,000$ to display the scalability of the model. Finally, the section is ended with a comparison study where we consider $G_4 = .67N(-3, \sqrt{2}^2) + .33N(3, \sqrt{2}^2)$, as in Martin and Tokdar (2012).

2.6.1 Simulation with G_1

For G_1, data were simulated as described above, and analyzed with $a_w = 1$, $b_w = 1$, $J = 5$, and c random. As seen in Figure 2.2 the density estimation is accurate for both (i) unknown, common and (ii) known variances, particularly considering that only 500 of the 3000 data points come from the non-null distribution. This results in the method's ability to keep both FDR and FNR relatively low, as seen in the summary of FDP and FNP in Table 2.2. The results from approximating Scott and Berger (2006) by setting $J = 8$ and fixing $c = 100,000$ are also contained in Table 2.2. The results are similar, but the parametric approach has a slightly larger number of rejections; this can be attributed to the parametric assumption of the model being met.

Table 2.2: Errors summary over one hundred simulations for G_1; MNR is the mean number of rejections.

	$J = 5$, c random						$J = 8$, $c = 10^5$					
	Common Variance			Known Variance			Common Variance			Known Variance		
Cost κ	MNR	FDR	FNR	MNR	FDR	FNR	MNR	FDR	FNR	MNR	FDR	FNR
1/5	234	0.024	0.098	248	0.025	0.093	238	0.027	0.097	249	0.025	0.093
1/3	250	0.038	0.094	264	0.039	0.090	255	0.042	0.093	265	0.040	0.090
1	301	0.099	0.085	315	0.106	0.081	309	0.109	0.083	316	0.107	0.081
3	399	0.227	0.074	422	0.247	0.070	414	0.247	0.072	424	0.249	0.070
5	480	0.318	0.069	517	0.347	0.065	503	0.341	0.067	519	0.348	0.065

2.6.2 Simulation with G_2

For G_2, data were simulated as described above, and analyzed. As seen in Figure 2.2 relatively good density estimation is still enjoyed for this skewed, median-zero mixture of two Gaussians. Improvement of the nonparametric Polya tree approach is clear; the density estimates of Scott and Berger (2006) miss the bimodal density and fit one Gaussian density over it.

The summary of FDP and FNP in Table 2.3 show the ability to control FDR and FNR is preserved. When comparing the nonparametric Polya tree approach to that of Scott and Berger (2006), beyond the novelty of this much improved density

estimate, it may be surprising that the FDR and FNR are similar across varied levels of cost κ; this is largely to do with the density estimates near zero being very similar in both approaches.

Table 2.3: Errors summary over one hundred simulations for G_2.

	$J = 5$, c random						$J = 8$, $c = 10^5$					
	Common Variance			Known Variance			Common Variance			Known Variance		
Cost κ	MNR	FDR	FNR	MNR	FDR	FNR	MNR	FDR	FNR	MNR	FDR	FNR
1/5	185	0.025	0.113	201	0.023	0.108	189	0.028	0.112	200	0.026	0.109
1/3	197	0.039	0.111	214	0.038	0.105	201	0.042	0.110	213	0.039	0.106
1	243	0.109	0.103	256	0.101	0.098	242	0.110	0.103	255	0.105	0.099
3	351	0.269	0.092	352	0.248	0.088	327	0.247	0.095	345	0.246	0.090
5	451	0.375	0.086	443	0.349	0.083	399	0.336	0.090	423	0.340	0.086

2.6.3 Simulation with G_3

For G_3, data were simulated as described above, and analyzed. As seen in Figure 2.2 good density estimation is still enjoyed for this symmetric, median-zero mixture of two Gaussians. The improvement of the nonparametric Polya tree approach again is clear; the density estimates of Scott and Berger (2006) miss the bimodal density and fit one Gaussian density over it.

The summary of FDP and FNP in Table 2.4 show the ability to control FDR and FNR is preserved. When comparing the nonparametric Polya tree approach to that of Scott and Berger (2006), beyond the novelty of this much improved density estimate, FDR and FNR are starkly different in this case across varied levels of cost κ. This improvement is due to the non-parametric approach being capable of capturing the non-Gaussian distribution of the non-null observations and, in this case, the estimates at zero are very different.

Table 2.4: Errors summary over one hundred simulations for G_3.

	$J = 5$, c random						$J = 8$, $c = 10^5$					
	Common Variance			Known Variance			Common Variance			Known Variance		
Cost κ	MNR	FDR	FNR	MNR	FDR	FNR	MNR	FDR	FNR	MNR	FDR	FNR
1/5	491	0.008	0.005	478	0.015	0.011	525	0.056	0.001	483	0.022	0.011
1/3	495	0.012	0.004	486	0.023	0.010	541	0.082	0.001	495	0.036	0.009
1	505	0.024	0.003	509	0.050	0.006	602	0.174	0.001	541	0.100	0.005
3	523	0.053	0.002	547	0.102	0.004	748	0.333	0.000	655	0.245	0.002
5	538	0.077	0.001	571	0.136	0.003	871	0.427	0.000	767	0.352	0.001

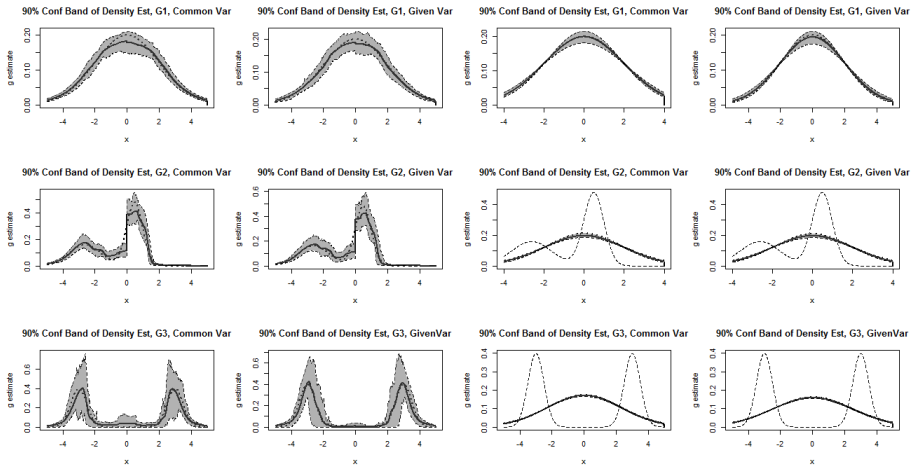

Figure 2.2: The first two columns show simulation results for G_1, G_2 and G_3 using the nonparametric approach and the last two columns show Scott and Berger (2006) results for the same data. The dotted densities are the true densities of the non-zero means and the solid density with the gray band is the estimated density with the 90 percent confidence band.

2.6.4 Scalability Results

For each G_j, $j = 1, 2, 3$, one hundred datasets were simulated with $\theta_1, \ldots, \theta_{5000} \overset{iid}{\sim} G_j$ and $\theta_i = 0$ for $i = 5001, \ldots, 30000$. Finally, we simulate (i) $y_i \overset{ind.}{\sim} N(\theta_i, 0.5^2)$ for the model with a common, unknown σ^2, or (ii) $y_i \overset{ind.}{\sim} N(\theta_i, \sigma_i^2)$ where $\sigma_i \sim \Gamma(5, 10)$ where $E(\sigma_i) = 0.5$. This nonparametric Polya tree method also retained the ability to keep both FDR and FNR relatively low with the expanded dataset. The density estimation is still very good for both common and known variance and with the expanded data set the density estimates are tighter in all three cases.

2.6.5 Simulation with G_1 Varying Prior Information

In this setting one hundred datasets were simulated for G_1,with $\theta_1, \ldots, \theta_{500} \overset{iid}{\sim} G_1$ and $\theta_i = 0$ for $i = 501, \ldots, 3000$. Finally, we simulate $y_i \overset{ind.}{\sim} N(\theta_i, \sigma_i^2)$ where $\sigma_i \sim \Gamma(5, 10)$ where $E(\sigma_i) = 0.5$. A variety of prior information $\hat{p}$, on w choosing $m = 10,000$ to ensure small variance was given and a summary of results are in Table 2.5. When perfect information $\hat{p} = 1/6$ on w is given the results are very similar

to the results of the default methodology with no prior information on w. As $\hat{p}$ is decreased, fewer observations are rejected which is expected as the prior information on w intimates that the model should be rejecting less. Recall Table 2.2 where decreasing the value of the relative cost, κ, exhibited similar behavior.

The most telling piece of this simulation are the density estimates seen in Figure 2.3; as the prior on w decreases, the mass around zero is slowly removed finally yielding "interestingly different" rejected observations and a density estimate of those observations.

Table 2.5: Errors summary over one hundred simulations for G_1 with known variance, $m = 10,000$, $\kappa = 1$, $J = 5$ and c random under T_{i1}; MNR is the mean number of rejections.

$\hat{p}$ Value	MNR	FDR	FNR
500/3000	316	0.107	0.081
315/3000	293	0.077	0.084
248/3000	284	0.066	0.086
100/3000	260	0.039	0.091

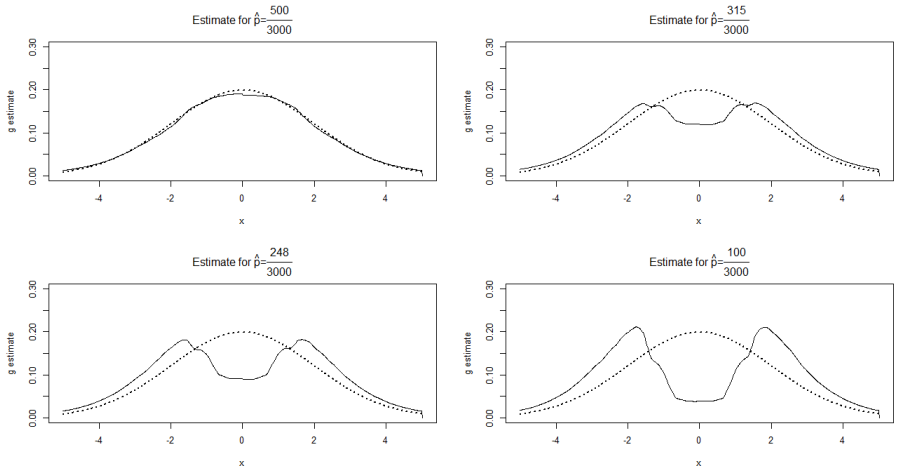

Figure 2.3: Density estimates over different values of $\hat{p}$ are the solid density estimates and density estimate with no prior on w is the dotted estimate.

2.6.6 Simulation with G_1 Varying Levels w

One hundred datasets were simulated for G_1,with $\theta_1, \ldots, \theta_{3000w} \overset{iid}{\sim} G_1$ and $\theta_i = 0$ for $i = 3000w + 1, \ldots, 3000$ for $w \in \{0.01, 0.\overline{33}, 0.1\overline{66}, 0.5, 0.8\overline{33}, 0.9\overline{66}, 0.99\}$. Finally,

we simulate $y_i \overset{ind.}{\sim} N(\theta_i, \sigma_i^2)$ where $\sigma_i \sim \Gamma(5, 10)$ where $E(\sigma_i) = 0.5$ for each value of w.

As seen in Table 2.6 the model does suffer in the "needle in a haystack" scenario. This is a direct result of the conditional distribution depending only on those observations for which $z_i > 0$, i.e. those observations deemed to be from the non-null distribution. When there are a small number of observations from the non-null distribution it is difficult for the Polya Tree approach to achieve decent distributional results. In the "hay in a needle stack" case this model keeps FDR low, but fails to discover many of the non-null observations. This is largely due to the non-null distribution having considerable mass at zero.

Table 2.6: Errors summary over one hundred simulations for G_1 with known variance, $\kappa = 1$, $J = 5$ and c random under T_{i1}.

	Common Variance			Known Variance		
w	MNR	FDR	FNR	MNR	FDR	FNR
30/3000	15	0.212	0.006	15	0.144	0.005
100/3000	52	0.129	0.018	54	0.120	0.018
500/3000	301	0.100	0.084	315	0.106	0.081
1500/3000	999	0.065	0.282	1127	0.098	0.257
2500/3000	1777	0.030	0.633	2319	0.082	0.543
2900/3000	1634	0.014	0.672	2906	0.029	0.792
2970/3000	1484	0.004	0.638	2964	0.010	0.917

2.6.7 Simulation with G_1 Varying Levels J

One hundred datasets for G_1 were simulated, with $\theta_1, \ldots, \theta_{500} \overset{iid}{\sim} G_1$ and $\theta_i = 0$ for $i = 501, \ldots, 3000$. Finally, we simulate $y_i \overset{ind.}{\sim} N(\theta_i, \sigma_i^2)$ where $\sigma_i \sim \Gamma(5, 10)$ where $E(\sigma_i) = 0.5$. This time the number of levels, J, in the Polya tree was varied to display its effect on the model. As seen in Table 2.7, FDR and FNR vary little over different levels of J. This shows robustness to the choice J; i.e. a "leveling off" noted by Hanson (2006).

Table 2.7: Errors summary over one hundred simulations for G_1 with known variance, $\kappa = 1$ and c random under T_{i1}.

Levels J Value	MNR	FDR	FNR
$J = 5$	315	0.106	0.081
$J = 6$	316	0.107	0.011
$J = 7$	317	0.108	0.081
$J = 8$	317	0.109	0.081

2.6.8 Simulation with G_1 and Wrongly Assumed Common Variance

One hundred datasets were simulated for G_1, with $\theta_1, \ldots, \theta_{500} \overset{iid}{\sim} G_1$ and $\theta_i = 0$ for $i = 501, \ldots, 3000$. Finally, we simulate $y_i \overset{ind.}{\sim} N(\theta_i, \sigma_i^2)$ where $\sigma_i \sim \Gamma(5, 10)$ where $E(\sigma_i) = 0.5$. We analyzed these data with the common variance approach to examine the models sensitivity to this assumption.

As seen in Table 2.8, FDR is much inflated compared to the results in Table 2.2 when we analyzed these simulated data with the correct, given variance approach. This indicates that the model is sensitive to a faulty assumption of common variance and shows the benefit of the extension of this model to the uncommon variance cases as discussed in Section 4.

Table 2.8: Errors summary over 100 simulations for G_1 with known variance analyzed under the common variance assumption for $J = 5$ and c fixed at $100,000$ under T_{i1}; MNR is the mean number of rejections.

Cost κ	MNR	FDR	FNR
1/5	446	0.324	0.078
1/3	490	0.359	0.075
1	625	0.449	0.067
3	846	0.551	0.058
5	1009	0.605	0.053

2.6.9 Comparison Study

For comparison, consider the simulation study of Martin and Tokdar (2012) who compare their predictive recursion (PRTest) to the Bayes oracle test and the methods of Jin and Cai (2007) and Muralidharan (2010) (mixfdr). Five hundred datasets for G_4 were simulated, with $\theta_i = 0$ for $\theta_1, \ldots, \theta_{1000(1-w)}$ and $\theta_i \overset{iid}{\sim} G_4$ for $i = 1000(1 -$

$w), \ldots, 1000$ for $w \in \{0.01, 0.05, 0.10, 0.15, 0.20, 0.25\}$. Finally, we simulate $y_i \overset{ind.}{\sim} N(\theta_i, \sigma^2)$ where $\sigma = 1$.

In Figure 2.4 it is seen that the Polya tree approach, with $\kappa = .01$, suffers high FDR in the sparse case where $w = .01$ and there are only ten observations from the non-null distribution. For the less sparse cases the model performs comparably to the other models. In these less sparse cases, e.g. $w \in \{.20, .25\}$, only the model of Jin and Cai (2007) has more power.

The flexibility of the model's cost based threshold is that we allow the user to give information about how costly a false discovery is relative to a false non-discovery. Note that in some scientific studies that the model with the lowest FDR is not always preferred. For instance, in an experiment where it is cheap to explore a discovery, but it is important that we find all of the non-null observations a higher FDR may be desired; this case is accommodated by setting κ large.

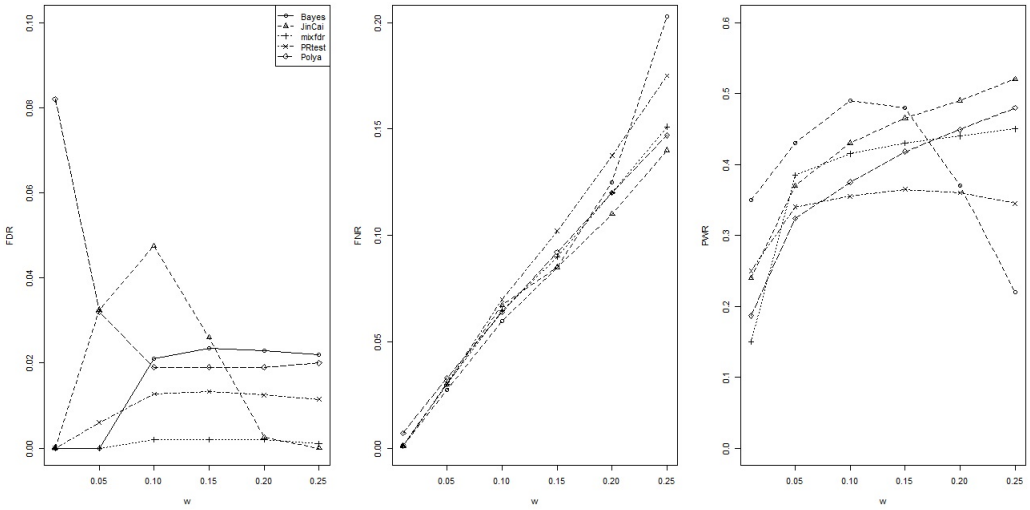

Figure 2.4: Plots of FDR (left), FNR (middle) and power (right) over varying values of w for G_4.

2.7 Data Analyses

Below this approach is applied to three different hypothesis testing scenarios: a difference of proportions, a paired difference, and a two sample difference.

2.7.1 Proportional Difference

Sun and McLain (2012) provide an example using educational survey data from the Adequate Yearly Progress study on the academic performances of students across different social and fiscal demographics; it is of interest to find schools where the difference is interestingly different from the norm. That is, schools that do significantly worse than other schools in serving their social-economically disadvantaged than social-economically advantaged students should be evaluated and efforts should be made to close the gaps. Along the same lines, schools that do significantly better can be analyzed to help educators find what helps close the gap.

Let $y_i = \hat{p}_{SA_i} - \hat{p}_{SD_i}$ where $\hat{p}_{SA_i}$ is the success rates of social-economically advantaged students in math exams at school i and $\hat{p}_{SD_i}$ is the success rate of social-economically disadvantaged students in math exams at school i with $\sigma_i^2 = \frac{\hat{p}_{SA_i}\left(1-\hat{p}_{SA_i}\right)}{n_{SA_i}} + \frac{\hat{p}_{SD_i}\left(1-\hat{p}_{SD_i}\right)}{n_{SD_i}}$ for $i = 1, \ldots, 7866$, as in the case of the usual Z statistic approach for proportions. In this particular setting, the students do differently on average so the observations are shifted to be median zero. Here, take $y_{shifted_i} = \hat{p}_{SA_i} - \hat{p}_{SD_i} - \eta$ where η is the median of all y_i's. It should be noted that the analyses here is directed at finding schools that are significantly different from other schools and not, in the traditional sense, finding schools that have dissimilar success rates for social-economically advantaged and disadvantaged students.'

The usual T-statistic approach yields 5,731 and 5,246 rejections at the 0.10 and 0.05 level of significance respectively. The default, no prior information on w, approach estimates $w = .9558$ with a 95% credible interval of the number of non-null

observations to be $(6801, 7851)$. The number of rejections from the model without using any prior information on w are 1863, 2298, 4468, 7511, and 7736 for costs of 1/5, 1/3, 1, 3, and 5 respectively. As cost, κ, is decreased the model rejects the null of fewer observations by design of threshold T_{i2} and with cost $\kappa = 1$ a similar number of rejections to that of the usual T-statistic approach at the 0.10 and .05 levels. The non-null density estimate, in this case, is approximately bell-shaped and symmetric at zero.

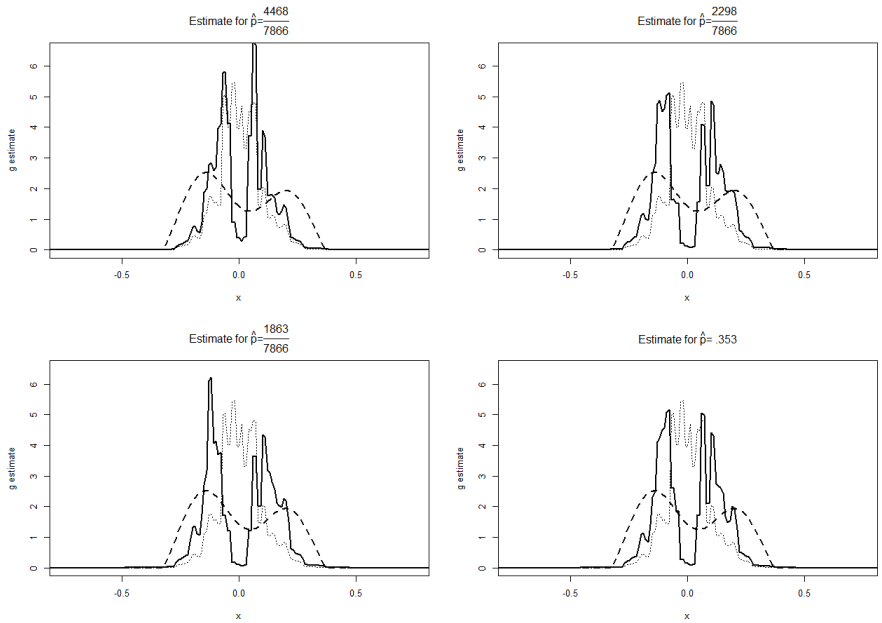

Figure 2.5: Density estimates over different values of $\hat{p}$ are the solid density estimates and density estimate with no prior on w is the dotted estimate. The density estimate from Sun and McLain (2012) is represented by the dashed density estimate.

Sun and McLain (2012) explore this data, searching for schools that are "interestingly different." We can consider this problem by inputting prior information on w. Sun and McLain (2012) consider the "oracle" method in Jin (2008) by using the $\hat{w} = .353$ which can be used as the prior information on w. Using this prior information the model reports a 95% credible interval of the number of non-null observations to be $(3366, 3545)$ and the non-null density estimate is bimodal indicating that there are two groups of "interestingly" different schools. In the absence of prior informa-

tion, or the use of an "oracle" method, similar results could be achieved by using the proportion of rejections in the default, no prior information on w, approach for small values of κ. As seen in Figure 2.5, using the proportions created by the number of rejections at $\kappa = 1$, $\kappa = 1/3$ and $\kappa = 1/10$ the model obtains results similar to Sun and McLain (2012).

2.7.2 Paired Difference

Next, consider the gene expression data from the microarray experiments of Colon tissue samples of Notterman et al. (2001). This data, available through the Princeton University Gene Expression Project, consists of 7,457 genes measurements for 18 patients on both tumor and normal tissues. The analysis here is directed at finding genes that are significantly different from other genes and not, in the traditional sense, finding genes that have changed expression at all. It is of importance to find the genes different than the other genes, in general; the genes that have significantly different expression changes should have their association with the disease of interest, colorectal adenoma here, further examined.

In the case of the usual, paired Student T statistic take $y_i = \bar{x}_{d_i}$ where $\bar{x}_{d_i}$ is the mean pairwise difference and $\sigma_i^2 = s_{d_i}^2/18$ for $i = 1, \ldots, 7457$. The data, in this example, are not median-zero so take shifted $y_{shifted_i} = \bar{x}_{d_i} - \eta$ where η is the median of all y_i's. Similar to the analysis of data from Sun and McLain (2012), the shift is necessary here as gene expression tends to change in congress (Morley et al., 2004). The usual T-statistic approach yields 2,818 and 2,205 rejections at the 0.10 and 0.05 level of significance respectively. The default approach, no prior information on w, estimates $w = .3621$ with a 95% credible interval of the number of non-null observations to be $(2577, 2839)$. The number of rejections from the model without using any prior information on w are 740, 824, 1029, 1346, and 1533 for costs of 1/5, 1/3, 1, 3, and 5 respectively. As we decrease the cost, κ, we see that the model

rejects the null of fewer observations by design of threshold T_{i1}. The non-null density estimate, in this case, is approximately bell-shaped and symmetric at zero.

2.7.3 Two Sample Difference

Ausin et al. (2011) consider the gene expression data from the microarray experiments of Colon tissue samples of Alon et al. (1999). This data, available through the R package "plsgenomics"consists of 2,000 genes for 62 samples; 40 of these samples are from tumor tissues and 22 are from normal tissues. It should be noted that the original dataset consisted of 6,500 genes and the 2,000 observations in the available data are the genes with highest minimal intensity.

In the case of the usual, two sample Student T statistic take $y_i = \bar{x}_{norm_i} - \bar{x}_{tumor_i}$ and $\sigma_i^2 = s^2_{norm_i}/22 + s^2_{tumor_i}/40$ for $i = 1, \ldots, 2,000$. Similar to the analyses of Notterman et al. (2001), take shifted $y_{shifted_i} = \bar{x}_{norm_i} - \bar{x}_{tumor_i} - \eta$ where η is the median of all y_i's. The usual T-statistic approach yields 626 and 477 rejections at the 0.10 and 0.05 level of significance respectively. Ausin et al. (2011) report 223 differentially expressed genes with posterior probability one. The default, no prior information on w, approach estimates $w = .7178$ with a 95% credible interval of $(1139, 1801)$. The number of rejections from the model without using any prior information on w are 126, 159, 240, 467, and 633 for costs of 1/5, 1/3, 1, 3, and 5 respectively.

2.8 Java Applet

An increasing trend among statisticians is to provide R packages for fitting complex methodology. These packages allow others familiar with the R computing environment to implement methods otherwise not readily available, and hence allow the routine use of new methodology. However, many scientists in other fields are unfamil-

iar with R and other programming languages. These scientists often use specialized software or Excel to implement statistical methods.

Perhaps the most widely available venue for the dissemination of new statistical methods is through the use of online Java applets that are compatible with any machine that can run Java; it is in this spirit that we provide such a Java applet. Scientists of all backgrounds are able to utilize this methodology without the steep learning curve of learning a new programming language. Figure 2.6 shows a screenshot of the Java applet with example input and output.

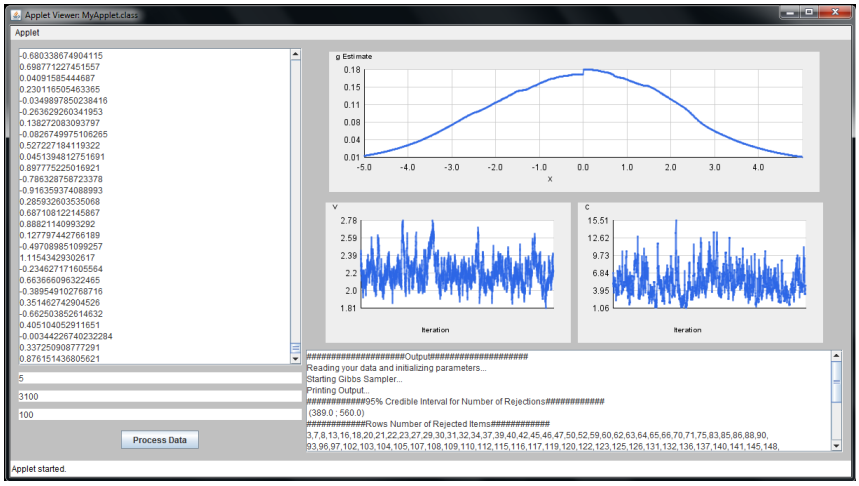

Figure 2.6: Screenshot of the Java application with results

The left side of the applet has text areas for input. The top text area is for entering the observations; each observation should have its own row. The applet is smart enough to know whether or not the input contains known variances and in the given variance case we expect the observation and variance to be comma separated; again, each couple should have its own row. In the second row, the two text areas are for the desired the number of Gibbs iterates and the number of burn in iterates, respectively. In the third row, the two text areas are for the desired cost κ and number of levels for the Polya tree J. The following line for input asks whether or not prior information on w will be used followed by two text areas - the estimated proportion of rejections and the constant m. The last line inquires whether the test

is a difference of means or differences of proportions so that the program can toggle between thresholds T_{i1} and T_{i2}.

The right side of the applet displays output. The first graph is the density estimate for the non-zero means, and the next two show the variables v and c after burn in. If the input is set correctly these graphs should appear fairly consistent over the iterates. If the graphs are not consistent, burn in and Gibbs iterates may need to be increased. The textual output gives a credible interval for the number of observations rejected and a list of which observations were rejected overall according to the threshold. This applet is available through http://people.stat.sc.edu/Cipolli/BMT/BMT.html.

2.9 Conclusion

The suggested approximate finite Polya tree multiple testing procedure is very successful in correctly classifying the observations with non-zero mean in a computationally efficient manner. This holds even when the non-zero means are simulated from a mean zero distribution, as seen in the simulation of θ_i from a $N(0, 2^2)$, which is particularly impressive as we can expect many of these 'non-zero' means to be very close to zero. The flexibility of this model is displayed through the data analyses in Section 7 by completing the task of multiple testing in the cases of proportional differences as well as paired and two sample mean differences.

The performance of the multiple comparisons was evaluated using FDR and FNR, both of which were kept low during simulation for relatively small and large numbers of observations. A nice aspect of the methodology and Java applet is that we are able to provide an approximation of the density of the non-zero means that is very close to the actual density function even in the non-Gaussian case. Further, the model is capable of this for "interestingly different" observations in the cases where that is of interest as in Section 6.5 and Section 7.1.

This model assumes that θ_i and σ_i^2 are independent and is sensitive to the prior

specifications including w. The model is also sensitive to the data being median zero; simulation results, not included, show that deviations from this lead to inflated error.

2.10 Acknowledgements

We would like to acknowledge and thank the editor and anonymous reviewers for providing insightful and constructive comments, suggestions and direction throughout revisions of this paper.

Chapter 3

Computationally Tractable Approximate and Smoothed Polya Tree

A discrete approximation to the Polya tree prior suitable for latent data is proposed that enjoys surprisingly simple and efficient conjugate updating. This approximation is illustrated in two applied contexts: the implementation of a nonparametric meta-analysis involving studies on the relationship between alcohol consumption and breast cancer, and random intercept Poisson regression for Ache armadillo hunting treks. The discrete approximation is then smoothed with Gaussian kernels to provide a smooth density for use with continuous data; the smoothed approximation is illustrated on a classic dataset on galaxy velocities and on recent data involving breast cancer survival in Louisiana.

3.1 Introduction

The field of Bayesian nonparametrics has exploded over the last two decades, largely due to several recent groundbreaking advances in computational techniques, in particular Markov chain Monte Carlo (MCMC). Every year sees an increase in papers advancing models, theory, and applications of Bayesian nonparametrics. Classical parametric statistics posits a family of probability models, say G, for data $\mathbf{y} = (y_1, \ldots, y_n)'$ indexed by a finite-dimensional parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$, e.g. normal-errors regression, generalized linear mixed models with Gaussian random effects, logistic regression, etc. A Bayesian nonparametric model seeks to generalize and robustify

posterior inference on functionals of the population distribution G by increasing the flexibility of the probability model, thus anticipating nonstandard, yet interesting aspects such as multimodality, extreme skew, nonlinear and irregular trends, and other features not effectively handled by many parametric models. This is often accomplished by defining a prior that is indexed by a very large number, typically infinite number of parameters. Furthermore, many Bayesian nonparametric priors generalize common parametric models such as Gaussian, linear trends in regression, etc., thus allowing testing and assessment of these simpler, but often parsimonious and interpretable models.

For the case of estimating a distribution function, Bayesian nonparametric approaches place a prior on the space of distribution functions; examples are the celebrated Dirichlet process, Polya tree priors, neutral to the right priors, Bernstein polynomials with random degree, transformed Gaussian processes, and many others. Some of these priors can place positive mass on distribution functions that admit a density with respect to Lebesgue measure (e.g. Polya trees, Bernstein polynomials, Dirichlet process mixtures of continuous kernels), and others do not (e.g. Dirichlet processes, neutral to the right priors). Ferguson (1973) effectively ushered in half a century of Bayesian nonparametrics with his foundational paper on the Dirichlet process, comprehensively treating common inferential problems such as the estimation of the population cumulative distribution function, mean, quantiles, variance, and covariance; as well as a treatment of the two-sample problem. So many important papers followed that it is impossible to provide a concise summary. Books providing an overview of Bayesian nonparametrics include Ghosh and Ramamoorthi (2003), Hjort et al. (2010), Mitra and Müller (2015), and Müller et al. (2015).

The Dirichlet process and Dirichlet process mixture (DPM) models have enjoyed tremendous success in the field of Bayesian nonparametrics over the last twenty years, yielding hundreds of applications papers, generalizations, and computational ad-

vances. Consider a Dirichlet process centered at the Gaussian distribution $G|c,\mu,\sigma \sim DP(c, N(\mu,\sigma^2))$. Under the Dirichlet process, or any stick-breaking process, the probability measure G is discrete, i.e. $G = \sum_{j=1}^{\infty} w_j \delta_{x_j}$ where δ_x is Dirac measure at x. Sethuraman (1994) showed for the Dirichlet process that $w_j = u_j \sum_{k=1}^{j-1}(1-u_k)$ for $u_1, u_2, \cdots \overset{iid}{\sim} \text{beta}(1,c)$ independent of $x_1, x_2, \cdots \overset{iid}{\sim} N(\mu,\sigma^2)$. The Dirichlet process was used as a latent random effects distribution in generalized linear mixed models (GLMMs) by Kleinman and Ibrahim (1998) and Jara et al. (2009), and as latent means in a meta-analysis by Burr and Doss (2005). Note that frequentists have been using finite discrete mixtures as random effects distributions in GLMMs for almost 30 years (Follman and Lambert, 1989; Aitkin, 1999).

This paper proposes two nonparametric priors based on Polya trees (Lavine, 1992), a prior over discrete probability measures suitable for random effects and a prior over continuous probability measures for density estimation and regression. Consider a finite Polya tree prior centered at $N(\mu,\sigma^2)$, $G|c,\mu,\sigma \sim PT_J(c, N(\mu,\sigma^2))$ (Hanson, 2006) where c intimates how much the Polya tree assimilates the centering distribution. The first prior, which can be used for latent unobservables such as random effects, is a discrete approximation to G that keeps the dyadic tree structure of Ferguson (1974) and Lavine (1992) on conditional probabilities, terminates the tree at some finite level J, and simply replaces the sets at level J with point mass. The second prior smooths this discrete approximation for use with continuous data. The resulting MCMC schemes enjoys conjugate updating for almost all parameters and excellent mixing, even for small c.

Section 3.2 introduces the approximate Polya tree (APT), which replaces the random G with a collection of point masses; considers the APT for modeling Gaussian means with known variances; implements the APT in random intercept generalized linear mixed models; and then proposes a smoothed APT (SAPT) suitable for continuous distributions but considerably more tractable than the usual mixture of Polya

trees (MPT) model. The SAPT is further generalized to the accelerated failure time model for censored survival data. All models are accompanied by simple MCMC schemes for posterior updating. Simulated and real examples are given in Section 3.3; Section 3.4 concludes the paper.

3.2 Models

3.2.1 Polya tree and discrete approximation

The Polya tree prior was introduced over several papers in the 1960's, summarized by Ferguson (1974). Polya trees were further developed by Lavine (1992, 1994) and Mauldin et al. (1992). Hanson (2006) discusses inference for mixtures of finite Polya trees, which smooth out the effect of the partition on posterior inference. Briefly, the finite Polya tree prior $G \sim PT_J(c, N(\mu, \sigma^2))$ adds to $N(\mu, \sigma^2)$ conditional probabilities that adjust the normal density's shape on intervals that partition $\mathbb{R}$; other centering families besides Gaussian can also be considered. The intervals are given by $B_{\mu,\sigma}(j,k) = (\mu + \sigma\Phi^{-1}\{(k-1)/2^j\}, \mu + \sigma\Phi^{-1}\{k/2^j\})$ where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function for a standard Gaussian; the 2^j intervals $B_{\mu,\sigma}(j,1), B_{\mu,\sigma}(j,2), \ldots, B_{\mu,\sigma}(j,2^j)$ partition $\mathbb{R}$ up to a set of Lebesgue measure zero. Let $Y_{j,2k-1}$ and $Y_{j,2k}$ be the G-probability of $B_{\mu,\sigma}(j,2k-1)$ and $B_{\mu,\sigma}(j,2k)$ given the event $B_{\mu,\sigma}(j-1,k)$; note that the first two sets partition the third. These conditional probabilities have prior

$$Y_{j,k}|c \overset{ind.}{\sim} \text{beta}(cj^2, cj^2), \tag{3.1}$$

where $j = 1, \ldots, J$ and k are the odd numbers from 1 to $2^j - 1$ at any level j. For any $Y_{j,k}$ where k is odd, let $Y_{j,k+1} = 1 - Y_{j,k}$. Define $\mathcal{Y} = \{Y_{j,k} : j = 1, \ldots, J; k = 1, \ldots, 2^j\}$. The G-probability of set k in the finest partition J, i.e. $G\{B_{\mu,\sigma}(J,k)\}$ is

$$p_{\mathcal{Y}}(k) = \prod_{j=1}^{J} Y_{j,\lceil k2^{j-J}\rceil}, \quad k = 1, \ldots, 2^J, \tag{3.2}$$

where $\lceil\cdot\rceil$ is the usual ceiling function. Then, for $G \sim PT_J(c, N(\mu, \sigma^2))$, the density $g(x|\mathcal{Y}, \mu, \sigma)$ of G given $\mathcal{Y}$, μ, and σ has the following form:

$$g(x|\mathcal{Y}, \mu, \sigma) = 2^J p_{\mathcal{Y}}\{k_{\mu,\sigma}(x)\}\phi(x|\mu, \sigma^2),$$

where 2^J gives the number of partitions in the last level of the Polya tree; $k_{\mu,\sigma}(x) = \lceil 2^J\Phi\{(x-\mu)/\sigma\}\rceil$ gives which set k at level J x is in, where $\phi(\cdot|\mu, \sigma^2)$ and $\Phi(\cdot|\mu, \sigma^2)$ are the density and cumulative distribution functions of a $N(\mu, \sigma^2)$ random variable,

respectively. Note that for any measurable $A \subset \mathbb{R}$, $E\{G(A|\mu,\sigma^2)\} = \int_A \phi(x|\mu,\sigma^2)dx$ where the expectation is over $\mathcal{Y}$ and c. This implies that the random measure G is centered at a $N(\mu,\sigma^2)$ distribution. The parameter $c > 0$ controls how "close" G is to $N(\mu,\sigma^2)$; as $c \to \infty$, $G(A) \to \int_A \phi(x|\mu,\sigma^2)dx$ almost surely. More details can be found in Hanson (2006).

Consider a discrete approximation to the Polya tree prior G, suitable for latent data or random effects, that simplifies computation enormously:

$$G(\cdot) = \sum_{k=1}^{2^J} p_{\mathcal{Y}}(k)\delta_{\mu+\sigma t_k}(\cdot), \quad t_k = \Phi^{-1}\left(\frac{k-0.5}{2^J}\right), \tag{3.3}$$

where $p_{\mathcal{Y}}(\cdot)$ is defined through (3.1) and (3.2). This can be a reasonable approximation; as J gets large, the sets $B_{\mu,\sigma}(J,k)$ get smaller, except in the tails, and $g(\cdot)$ varies less over the sets. Since $g(\cdot)$ follows $N(\mu,\sigma^2)$ on these sets, $g(\cdot)$ over the set can be approximated with just one "representative" point, the mid-quantile, plus the associated probability $p_{\mathcal{Y}}(k)$ of the set under G. This discrete approximation has one unappealing property: unlike the Dirichlet process, G does not have full weak support for $J < \infty$. For random effects distributions, we simply seek G to be flexible. Most existing approaches to modeling a random effects distribution G do not have full weak support, including Ghidey et al. (2004), Komárek and Lesaffre (2008), Jara et al. (2009), and many others.

The number of free parameters in $\mathcal{Y}$ is $2^J - 1$. As J increases G can accommodate finer and finer detail. Hanson (2006), however, noticed a "leveling off" effect in the log-pseudo marginal likelihood (LPML) (Gelfand and Dey, 1994), a cross-validated predictive measure of fit, across a wide variety of models using MPTs; i.e. after a certain point, increasing J does not enhance the predictive ability of a model. Hanson and Johnson (2002) suggest $J = \lceil \log_2 n \rceil$ and Hanson (2006) suggests $J = \lceil \log_2 n/N \rceil$ where $N = 5$ or $N = 10$ is roughly the number of observations informing G on the finest partition at J as rough guidelines, but in practice $J = 5$ or $J = 6$ has provided essentially identical inference as larger values of J across many datasets.

An unpublished Technical Report (available from the second author) shows how a prior can be placed on J for a marginalized Polya tree. This approach, however, cannot be used directly for the models in this paper. Reversible jump (Green, 1995) could offer a solution for the models presented here, but is not considered further.

Given (μ, σ^2), $E\{G[(-\infty, x)]\} \overset{J\to\infty}{\to} \Phi\{(x-\mu)/\sigma\}$, thus hyperpriors for (μ, σ) can be chosen as one would under the normal model. Section 5.2.3 in Christensen et al. (2010) provides guidance towards choosing an informative prior; alternatively Jeffreys' prior under the centering Gaussian model could be used, or simply a flat prior. The latter choice is especially easy to implement for the models considered in Section 3, and so the non-informative prior

$$p(\mu, \sigma) \propto I\{\sigma > 0\}, \tag{3.4}$$

is assumed. Call the prior (3.1), (3.2), (3.3), and (3.4) an approximate Gaussian-centered Polya tree prior of level J, denoted

$$G \sim APT_J.$$

The dependence of G on c is suppressed for now. We now consider several special cases, illustrating how MCMC simplifies.

3.2.2 Gaussian data with known variances

Assume the hierarchical model

$$y_i | x_i \sim N(x_i, s_i^2),$$

where the s_i are known, and

$$x_1, \ldots, x_n | G \overset{iid}{\sim} G, \quad G \sim APT_J.$$

This is an approximation to the model considered by Branscum and Hanson (2008) for the purpose of carrying out a nonparametric meta-analysis. Posterior updating is simplified by introducing latent q_i associated with each x_i; $q_i = k \Leftrightarrow x_i = \mu + \sigma t_k$ for $k = 1, \ldots, 2^J$. Let $\mathbf{y} = (y_1, \ldots, y_n)'$, $\mathbf{q} = (q_1, \ldots, q_n)'$, and $\setminus = (s_1, \ldots, s_n)'$. The likelihood augmented with $\mathbf{q}$ is

$$p(\mathbf{y}|\mu, \sigma, \mathbf{q}, \mathcal{Y}) = p(\mathbf{y}|\mu, \sigma, \mathbf{q}) p(\mathbf{q}|\mathcal{Y}) = \left[\prod_{i=1}^{n} \phi(y_i | \mu + \sigma t_{q_i}, s_i^2)\right] \left[\prod_{i=1}^{n} p_{\mathcal{Y}}(q_i)\right].$$

Note that the first portion is a simple weighted linear regression on the "intercept" μ and "slope" σ. This likelihood times the prior $p(\mu, \sigma)p(\mathcal{Y})$ yields the posterior density.

All parameters have closed-form full conditional distributions, and so a Gibbs sampler for sampling $(\mu, \sigma, \mathcal{Y}|\mathbf{y})$ is immediate. The full conditional distribution for each index $q_1, \ldots, q_n$ is

$$P(q_i = k|\mu, \sigma, \mathcal{Y}, y_i) \propto \phi(y_i|\mu + \sigma t_k, s_i^2) p_{\mathcal{Y}}(k), \quad k = 1, \ldots, 2^J. \tag{3.5}$$

Like all finite mixtures, including finite approximations to the Dirichlet process (Ishwaran and Zarepour, 2002), the computational burden grows with the number of components 2^J. However, one can update by proposing only a small subset of $\{1, \ldots, 2^J\}$, namely one of its immediate neighbors. This reduces the number of Gaussian density evaluations at each MCMC iteration from 2^J to 2, increasing speed enormously with only a minimal cost in terms of posterior mixing. If the previous value is $q_i = 1$, then $q_i^* = 2$; if $q_i = 2^J$ then $q_i^* = 2^J - 1$. If $2 \leq q_i \leq 2^J - 1$ then choose $q_i^* = q_i - 1$ or $q_i^* = q_i + 1$ with equal probability $\frac{1}{2}$. The proposed q_i^* is accepted with probability

$$1 \wedge \frac{\phi(y_i|\mu + \sigma t_{q_i^*}, s_i^2) p_{\mathcal{Y}}(q_i^*)}{\phi(y_i|\mu + \sigma t_{q_i}, s_i^2) p_{\mathcal{Y}}(q_i)} \frac{2^{I\{q_i=2\} I\{q_i^*=1\}} 2^{I\{q_i=2^J-1\} I\{q_i^*=2^J\}}}{2^{I\{q_i=1\}} 2^{I\{q_i=2^J\}}},$$

where $\wedge$ is "the smaller of." The resulting transition kernel can be verified as aperiodic, irreducible, and positive recurrent as $p_{\mathcal{Y}}(k) > 0$ almost surely for $k = 1, \ldots, 2^J$. The remaining transition kernels $[\mathcal{Y}|\mathbf{q}]$ and $[\mu, \sigma|\mathbf{q}]$ (below) are simple Gibbs and Metropolis-Hastings (M-H) updates, respectively; therefore the full chain obtained from the product of all transition kernels is easily seen to be ergodic.

Let $n(J, k) = \sum_{i=1}^n I\{q_i = k\}$ for $k = 1, \ldots, 2^J$. Then recursively, $n(j-1, k) = n(j, 2k-1) + n(j, 2k)$ for j going from J to 1. Then

$$Y_{j,2k-1}|\mathbf{q} \sim \text{beta}(cj^2 + n(j, 2k-1), cj^2 + n(j, 2k)),$$

for $k = 1, \ldots, 2^{j-1}$ and $j = 2, \ldots, J$. Here and elsewhere we fix $Y_{1,1} = Y_{1,2} = 0.5$ forcing $G(\mu) = 0.5$, i.e. μ is a median of G. Given $\mathbf{q}$, the weighted linear regression form of the likelihood implies that

$$\begin{bmatrix} \mu \\ \sigma \end{bmatrix} |\mathbf{q}, \mathbf{y} \sim N_2 \left(\mathbf{M} \begin{bmatrix} \sum_{i=1}^n \frac{y_i}{s_i^2} \\ \sum_{i=1}^n \frac{t_{q_i} y_i}{s_i^2} \end{bmatrix}, \mathbf{M} \right), \ \mathbf{M}^{-1} = \begin{bmatrix} \sum_{i=1}^n \frac{1}{s_i^2} & \sum_{i=1}^n \frac{t_{q_i}}{s_i^2} \\ \sum_{i=1}^n \frac{t_{q_i}}{s_i^2} & \sum_{i=1}^n \frac{t_{q_i}^2}{s_i^2} \end{bmatrix}. \tag{3.6}$$

If an informative prior is placed on (μ, σ), a candidate (μ^*, σ^*) can be proposed from this bivariate Gaussian full conditional distribution (under a flat prior), but accepted with M-H probability $1 \wedge p(\mu^*, \sigma^*)/p(\mu, \sigma)$. For prior (3.4) the proposal is simply accepted if σ^* is positive.

The probabilities $\mathcal{Y}$ are "unhinged" from the location of the sets, or points $\mu+\sigma t_k$; this allows for particularly simple updating relative to a standard finite mixture of Polya trees prior, which can have problematic updating and mixing for (μ, σ) (Hanson, 2006). In many, essentially all, papers involving MPT priors to date, updating (μ, σ) is not conjugate; updating proceeds typically via slice sampling, adaptive Metropolis, or random-walk Metropolis steps. Furthermore, the posterior mixing for these parameters is often terrible if the true distribution that G is modeling is very unlike the centering distribution $N(\mu, \sigma^2)$. As we shall see in Section 3.3, mixing is much better for the APT, and updating (μ, σ) through (3.6) is remarkably easy.

The parameter c can also be random; the full conditional distribution is the product of $2^J - 2$ beta densities:

$$p(c|\mathcal{Y}) \propto \prod_{j=2}^{J} \prod_{k=1}^{2^{j-1}} \frac{\Gamma(2cj^2)}{\Gamma(cj^2)^2} Y_{j,2k-1}^{cj^2-1}(1 - Y_{j,2k})^{cj^2-1}.$$

A beta random variable can be approximated by a logit-normal (Aitchison and Shen, 1980). Specifically, beta(a, a) is well-approximated by logit–$N(0, \frac{2}{a})$ (Zhao and Hanson, 2011). Thus $\log \frac{Y_{j,k}}{1-Y_{j,k}} \overset{\bullet}{\sim} N(0, \frac{2}{cj^2})$, where $\overset{\bullet}{\sim}$ means "approximately distributed as." This implies $p(\log \frac{Y_{j,k}}{1-Y_{j,k}}|c) \overset{\bullet}{\propto} c^{1/2} \exp\{-\frac{cj^2}{4}(\log \frac{Y_{j,k}}{1-Y_{j,k}})^2\}$, and finally $p(\{\log \frac{Y_{j,k}}{1-Y_{j,k}}\}|c) \overset{\bullet}{\propto} \prod_{j=2}^{J} \prod_{k=1}^{2^{j-1}} c^{1/2} \exp\{-\frac{cj^2}{4}(\log \frac{Y_{j,k}}{1-Y_{j,k}})^2\}$. This yields a gamma proposal

$$c^*|\mathcal{Y} \sim \Gamma\left(a_c + \frac{2^J - 2}{2}, b_c + \frac{1}{4}\sum_{j=2}^{2^J} j^2 \sum_{k=1}^{2^{j-1}} (\log \tfrac{Y_{j,2k-1}}{1-Y_{j,2k-1}})^2\right),$$

assuming $c \sim \Gamma(a_c, b_c)$. The proposal c^* is accepted with the usual M-H probability. Alternatively c can be updated with adaptive or random-walk Metropolis steps.

3.2.3 Random intercept GLMMs

Consider a simple univariate random-intercept GLMM with the APT_J prior on the distribution G of univariate random effects. Each observation y_i is accompanied by covariates $\mathbf{z}_i \in \mathbb{R}^p$. For example, conditional on $(\boldsymbol{\gamma}, x_i)$ the Poisson model with log link stipulates

$$y_i \sim \text{Pois}(N_i\lambda_i), \quad \log(\lambda_i) = \mathbf{z}_i'\boldsymbol{\gamma} + x_i,$$

and the binomial model with logit link is

$$y_i \sim \text{bin}(m_i, \pi_i), \quad \text{logit}(\pi_i) = \mathbf{z}_i'\boldsymbol{\gamma} + x_i.$$

In what follows, replace x_i by σt_{q_i}. The intercept μ is subsumed into $\boldsymbol{\gamma}$ yielding an augmented likelihood

$$\mathcal{L}(\boldsymbol{\gamma}, \sigma, \mathbf{q}) = \prod_{i=1}^{n} L_i(\boldsymbol{\gamma}, \sigma, q_i) = \prod_{i=1}^{n} \exp\{-N_i \exp(\mathbf{z}_i'\boldsymbol{\gamma} + \sigma t_{q_i})\} \exp\{(\mathbf{z}_i'\boldsymbol{\gamma} + \sigma t_{q_i}) y_i\},$$

for Poisson data, and

$$\mathcal{L}(\boldsymbol{\gamma}, \sigma, \mathbf{q}) = \prod_{i=1}^{n} L_i(\boldsymbol{\gamma}, \sigma, q_i) = \prod_{i=1}^{n} \frac{\exp\{(\mathbf{z}_i'\boldsymbol{\gamma} + \sigma t_{q_i}) y_i\}}{[1 + \exp(\mathbf{z}_i'\boldsymbol{\gamma} + \sigma t_{q_i})]^{m_i}}.$$

for binomial data. The latent q_i are updated similarly to (3.5)

$$P(q_i = k | \boldsymbol{\gamma}, \sigma, \mathcal{Y}, y_i) \propto L_i(\boldsymbol{\gamma}, \sigma, k) p_{\mathcal{Y}}(k), \quad k = 1, \ldots, 2^J.$$

The sampling of $(\boldsymbol{\beta}, \sigma)$ proceeds as in a standard GLM *without random effects*, and is easily accomplished via adaptive M-H or the approach of Gamerman (1997). Gamerman's approach is based on one iteration of a Newton-Raphson optimization scheme, i.e. iteratively reweighted least squares. It is similar in spirit to Langevin-adjusted M-H in that the gradient of the posterior density at the current value is used. This approach is easy to carry out and is described below for Poisson and binomial GLMMs. Let $\theta_i = E(y_i | \boldsymbol{\gamma}, \sigma, q_i)$, the conditional mean. For Poisson this is $\theta_i = N_i \lambda_i = N_i e^{\mathbf{z}_i'\gamma + \sigma t_{q_i}}$ and for binomial $\theta_i = m_i \pi_i = m_i \frac{e^{\mathbf{z}_i'\gamma + \sigma t_{q_i}}}{1 + e^{\mathbf{z}_i'\gamma + \sigma t_{q_i}}}$. Let $\mathbf{W}(\boldsymbol{\gamma}, \sigma, \mathbf{q}) = \text{diag}(w_{11}, \ldots, w_{nn})$ be an $n \times n$ diagonal matrix with the conditional variances, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$ be the vector of conditional means. For Poisson $w_{ii} = \theta_i$ and for binomial $w_{ii} = m_i \pi_i (1 - \pi_i)$. Assume the prior $(\boldsymbol{\gamma}, \sigma) \sim N_{p+1}(\boldsymbol{\gamma}_0, \mathbf{S})$. Let $\mathbf{Z}(\mathbf{q})$ be the $n \times (p+1)$ matrix with ith row $[\ \mathbf{z}_i' \quad t_{q_i}\ ]$. Let

$$\begin{aligned}
\mathbf{V}(\boldsymbol{\gamma}, \sigma, \mathbf{q}) &= [\mathbf{S}^{-1} + \mathbf{Z}(\mathbf{q})' \mathbf{W}(\boldsymbol{\gamma}, \sigma, \mathbf{q}) \mathbf{Z}(\mathbf{q})]^{-1}, \\
\mathbf{m}(\boldsymbol{\gamma}, \sigma, \mathbf{q}) &= \mathbf{V}(\boldsymbol{\gamma}, \sigma, \mathbf{q}) [\mathbf{S}^{-1} \boldsymbol{\gamma}_0 + \mathbf{Z}(\mathbf{q})' \mathbf{W}(\boldsymbol{\gamma}, \sigma, \mathbf{q}) \mathbf{Z}(\mathbf{q}) \boldsymbol{\gamma} + \mathbf{Z}(\mathbf{q})' (\mathbf{y} - \boldsymbol{\theta}(\boldsymbol{\gamma}, \sigma, \mathbf{q}))].
\end{aligned}$$

The proposal $(\boldsymbol{\gamma}^*, \sigma^*) \sim N_{p+1}(\mathbf{m}(\boldsymbol{\gamma}, \sigma, \mathbf{q}), \mathbf{V}(\boldsymbol{\gamma}, \sigma, \mathbf{q}))$ is accepted with probability

$$1 \wedge \frac{\phi_{p+1}(\boldsymbol{\gamma}, \sigma | \mathbf{m}(\boldsymbol{\gamma}^*, \sigma^*, \mathbf{q}), \mathbf{V}(\boldsymbol{\gamma}^*, \sigma^*, \mathbf{q})) \phi_{p+1}(\boldsymbol{\gamma}^* | \boldsymbol{\gamma}_0, \mathbf{S}) \mathcal{L}(\boldsymbol{\gamma}^*, \sigma^*, \mathbf{q})}{\phi_{p+1}(\boldsymbol{\gamma}^*, \sigma^* | \mathbf{m}(\boldsymbol{\gamma}, \sigma, \mathbf{q}), \mathbf{V}(\boldsymbol{\gamma}, \sigma, \mathbf{q})) \phi_{p+1}(\boldsymbol{\gamma} | \boldsymbol{\gamma}_0, \mathbf{S}) \mathcal{L}(\boldsymbol{\gamma}, \sigma, \mathbf{q})}.$$

3.2.4 Smoothed version for density estimation

The discrete approximation (3.3) is suitable for latent unobservables $x_1, \ldots, x_n$. However, for continuous data the process needs to be smoothed. We propose replacing the sets $B_{\mu,\sigma}(J, k)$ at level J with Gaussian kernels of approximately the same spread. Gaussian kernels are especially tractable as the (μ, σ) updates enjoy the same easy conjugacy of Section 3.2.2. The resulting density model has similarities to the penalized B-spline approximation of Komárek et al. (2005) on a fixed set of

knots, but with random degree α (introduced below) and a Polya tree "penalty function" that encourages multiresolution, wavelet-like shrinkage (Draper, 1999). The model is also similar in spirit to the convolution approach to approximating stochastic processes (Higdon, 2002). Each point $x_i = \mu + \sigma t_{q_i}$ is smoothed with a Gaussian kernel; the kernel standard deviation is proportional to distance between subsequent knots with proportionality constant $\alpha > 0$. Define the distances $d_i = t_i - t_{i-1}$ for $i = 2^{J-1} + 1, \ldots, 2^J$ and $d_i = t_{i+1} - t_i$ for $i = 1, \ldots, 2^{J-1}$. The smoothed density is

$$g(x|\mu, \sigma, \mathcal{Y}, \alpha) = \sum_{k=1}^{2^J} p_{\mathcal{Y}}(k)\phi(x|\mu + \sigma t_k, \alpha^2 d_k^2).$$

This is a similar framework as Section 3.2.2; the MCMC updates for q_i in (3.5) are the same setting $s_i = \alpha d_i$. The parameter α, along with c, is a smoothing parameter. Larger values of α increase the Gaussian kernel variance relative to the mass locations; this has a similar effect to increasing the degree of a B-spline with a fixed number of knots. The value of α should not be much smaller than σ, however, as this produces "spikes" of probability instead of a smooth density. Setting $\alpha = \sigma$ coupled with $p_{\mathcal{Y}}(k) = 2^{-J}$, i.e. the expected value under the prior, gives a density that closely resembles a $N(\mu, \sigma^2)$; see Figure 3.1. With this in mind, $\alpha^{-2}|\sigma \sim \Gamma(a, a\sigma^2)$ ensures that $\alpha \approx \sigma$ for large a ($a = 100$ has worked very well across a variety of datasets) and yields the full conditional distribution

$$\alpha^{-2}|\mathbf{q}, \mu, \sigma \sim \Gamma\left(a + \frac{n}{2}, a\sigma^2 + \frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \mu - \sigma t_{q_i})^2}{d_{q_i}^2}\right). \tag{3.7}$$

The update for (μ, σ) is similar to (3.6), except that the lower right element of $\mathbf{M}^{-1}$ has $\frac{2a}{\alpha^2}$ added to it and requires a M-H step, i.e. a proposal (μ^*, σ^*) is generated

$$\begin{bmatrix} \mu^* \\ \sigma^* \end{bmatrix} |\mathbf{q}, \alpha, \mathbf{y} \sim N_2\left(\mathbf{M}\begin{bmatrix} \sum_{i=1}^n \frac{y_i}{\alpha^2 d_{q_i}^2} \\ \sum_{i=1}^n \frac{t_{q_i} y_i}{\alpha^2 d_{q_i}^2} \end{bmatrix}, \mathbf{M}\right), \ \mathbf{M}^{-1} = \begin{bmatrix} \sum_{i=1}^n \frac{1}{\alpha^2 d_{q_i}^2} & \sum_{i=1}^n \frac{t_{q_i}}{\alpha^2 d_{q_i}^2} \\ \sum_{i=1}^n \frac{t_{q_i}}{\alpha^2 d_{q_i}^2} & \frac{2a}{\alpha^2} + \sum_{i=1}^n \frac{t_{q_i}^2}{\alpha^2 d_{q_i}^2} \end{bmatrix},$$

where (μ^*, σ^*) is accepted with probability $1 \wedge (\sigma^*/\sigma)^{2a}[p(\mu^*, \sigma^*)/p(\mu, \sigma)]$. The ability of the smoothed version to capture the Gaussian density hints at something deeper. Unser et al. (1992) show that a properly normalized B-spline converges to a Gaussian density. Gans and Gill (1984) show empirically that Gaussians can be captured by B-splines with only a handful (e.g. 5 or 7) of basis functions.

Canale and Dunson (2016) consider a related approach that follows Chen et al. (2014) by transforming the data to $[0, 1]$ through the centering measure and using a Bernstein polynomial. But rather than smooth the Polya tree only at level J as

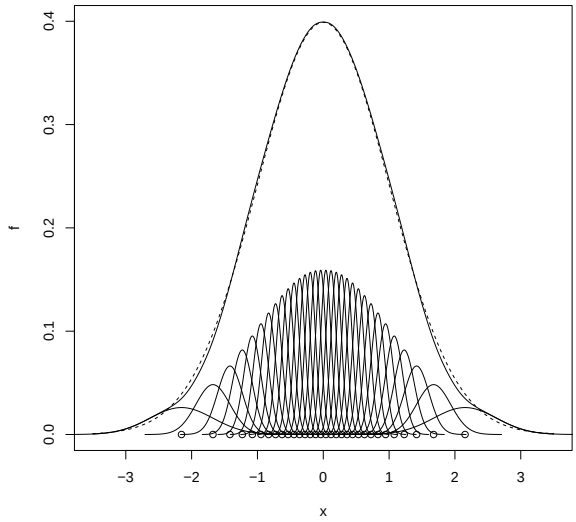

Figure 3.1: The 32 smoothing kernels for $J = 5$, a standard Gaussian density (dashed), and $g(x|0, 1, \mathcal{Y}, 1)$ with $Y_{j,k} = 0.5$. (solid)

presented here, Canale and Dunson (2016) consider many basis functions of different degrees, i.e. Bernstein polynomial basis functions from all degrees 2^j for integer j, and achieve sparsity through a stopping rule in the Polya tree in a manner similar to Wong and Ma (2010).

3.2.5 SAPT for regression error

A referee has suggested extending the SAPT of Section 3.2.4 to regression data $\{(\mathbf{z}_i, y_i)\}_{i=1}^n$, where $\mathbf{z}_i \in \mathbb{R}^p$, so that the location changes smoothly with covariates:

$$g_{\mathbf{z}}(x|\mu, \sigma, \mathcal{Y}, \alpha) = \sum_{k=1}^{2^J} p_{\mathcal{Y}}(k)\phi(x|\mathbf{z}'\boldsymbol{\gamma} + \sigma t_k, \alpha^2 d_k^2).$$

If each y_i is a log-survival time then this is the accelerated failure time model, e.g. Hanson and Johnson (2002), and e^{γ_j} is how the mean, median, or any quantile of survival changes when the jth predictor is increased by one unit. This changes posterior updating negligibly. The $(p+1)$-dimensional $(\boldsymbol{\gamma}, \sigma)$ has proposal

$$\begin{bmatrix} \gamma^* \\ \sigma^* \end{bmatrix} |\mathbf{q}, \alpha, \mathbf{y} \sim N_{p+1}\left(\mathbf{M}\begin{bmatrix} \sum_{i=1}^n \frac{y_i \mathbf{z}_i}{\alpha^2 d_{q_i}^2} \\ \sum_{i=1}^n \frac{t_{q_i} y_i}{\alpha^2 d_{q_i}^2} \end{bmatrix}, \mathbf{M}\right), \ \mathbf{M}^{-1} = \begin{bmatrix} \sum_{i=1}^n \frac{\mathbf{z}_i \mathbf{z}_i'}{d_{q_i}\alpha^2} & \sum_{i=1}^n \frac{\mathbf{z}_i t_{q_i}}{d_{q_i}\alpha^2} \\ \sum_{i=1}^n \frac{t_{q_i}\mathbf{z}_i'}{d_{q_i}\alpha^2} & \frac{2a}{\alpha^2} + \sum_{i=1}^n \frac{t_{q_i}^2}{d_{q_i}\alpha^2} \end{bmatrix},$$

accepted with probability $1 \wedge (\sigma^*/\sigma)^{2a}[p(\boldsymbol{\gamma}^*, \sigma^*)/p(\boldsymbol{\gamma}, \sigma)]$. Each index $q_1, \ldots, q_n$ has full conditional distribution

$$P(q_i = k|\mu, \sigma, \mathcal{Y}, y_i) \propto \phi(y_i|\mathbf{z}_i'\boldsymbol{\gamma} + \sigma t_k, s_i^2) p_{\mathcal{Y}}(k), \quad k = 1, \ldots, 2^J.$$

Finally,

$$\alpha^{-2}|\mathbf{q}, \boldsymbol{\gamma}, \sigma, \mathbf{y} \sim \Gamma\left(a + \frac{n}{2}, a\sigma^2 + \frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - \mathbf{z}_i'\boldsymbol{\gamma} - \sigma t_{q_i})^2}{d_{q_i}^2}\right).$$

Assume now survival data $\{(u_i, \mathbf{z}_i, \delta_i)\}_{i=1}^n$ where $\delta_i = 0$ indicates that the true survival time is greater than u_i, otherwise $\delta_i = 1$ indicates the survival time is u_i. Let $y_i = \log u_i$ for those i such that $\delta_i = 1$. The MCMC scheme can be augmented to update latent $[y_i|\boldsymbol{\gamma}, \sigma, \mathcal{Y}, \alpha, \mathbf{y}_{-i}]$ for $\delta_i = 0$, where $\mathbf{y}_{-i}$ is $\{y_j : j \neq i\}$, as $y_i \sim N(\mathbf{z}_i'\boldsymbol{\gamma} + \sigma t_{q_i}, \alpha^2 d_{q_i}^2)$ truncated to $y_i \in (\log u_i, \infty)$. The survival curve for covariates $\mathbf{z}$ is given by

$$S_{\mathbf{z}}(t|\mu, \sigma, \mathcal{Y}, \alpha) = 1 - \sum_{k=1}^{2^J} p_{\mathcal{Y}}(k) \Phi(\log t|\mathbf{z}'\boldsymbol{\gamma} + \sigma t_k, \alpha^2 d_k^2).$$

3.3 Illustrations

3.3.1 Alcohol and breast cancer risk

Following Branscum and Hanson (2008), we consider the meta-analysis data of Longnecker (1994). These data are comprised of $n = 39$ epidemiological studies aimed at exploring the relationship between alcohol consumption and risk of breast cancer in women. The data are the log-odds ratio y_i and standard error s_i from each study i. Specifically, the summary measure used was the estimated change in log odds ratio (scaled as $LOR \times 1000$) for a 1 gram increase in daily alcohol consumption. For example, Longnecker (1994) estimated that there is approximately a 10 percent increase of the risk of breast cancer in women for each 10-g/day increment in alcohol consumption. The model is $y_i|x_i \sim N(x_i, s_i^2)$, $x_1, \ldots, x_{39}|G \sim G$, $G|c \sim APT_J$ where $J = 5$ and $c \sim \Gamma(5, 1)$; the APT model is compared to MPT and Dirichlet process models for these data via the `PTmeta` and `DPmeta` functions in `DPpackage` (Jara et al., 2011) for R (R Core Team, 2014); all three models use an approximation to Jeffrey's prior on (μ, σ^2) for the underlying $N(\mu, \sigma^2)$ family. Specifically, the MPT model

assumes $G|c \sim PT_J(c, N(\mu, \sigma^2))$ where $\mu \sim N(0, 1000)$, $\sigma^{-2} \sim \Gamma(0.01, 0.01)$, and $c \sim \Gamma(5, 1)$. The Dirichlet process model assumes $G|c \sim DP(c, N(\mu, \sigma^2))$ with the same priors on (μ, σ) but takes $c \sim \Gamma(1, 1)$. See the `DPpackage` documentation for details on the model and these hyperparameters. In all three scenarios, a burn-in of $20,000$ was used and $10,000$ iterates were thinned from $1,000,000$.

For the APT model the posterior median of μ is 10.0 with a 95% credible interval of $(6.2, 14.2)$. These overall inferences on the log odds of breast cancer are plotted as vertical lines in Figure 3.2 along with the 39 study-specific log odds ratios $\times 1000$ and their 95% confidence intervals.

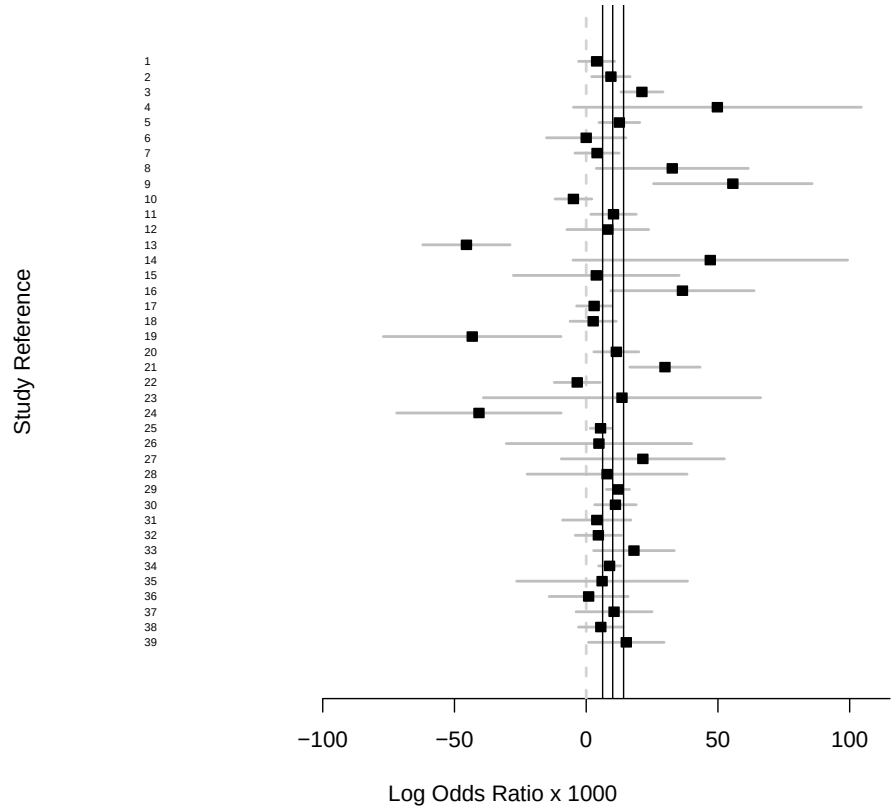

Figure 3.2: Log odds ratios with 95% CIs for the 39 studies. Vertical lines are posterior median and 95% CI for μ from the APT model.

For data $y_1, \ldots, y_n$ generated from a probability model $p(y_1, \ldots, y_n|\boldsymbol{\theta})$ indexed by

$\boldsymbol{\theta} \in \boldsymbol{\Theta}$ with prior $\boldsymbol{\theta} \sim q(\boldsymbol{\theta})$, the conditional predictive ordinate (CPO) for the ith observation y_i is

$$\text{CPO}_i = p(y_i|\mathbf{y}_{-i}) = \int_{\boldsymbol{\Theta}} p(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}_{-i}) d\boldsymbol{\theta}.$$

This is predictive distribution for a new observation given $\mathbf{y}_{-i} = \{y_j\}_{j \neq i}$ evaluated at the value that was left out y_i; as usual $p(\boldsymbol{\theta}|\mathbf{y}_{-i}) \propto p(\mathbf{y}_{-i}|\boldsymbol{\theta}) q(\boldsymbol{\theta})$. The larger CPO_i is, the better supported y_i is through the remaining data, model, and prior. (Gelfand and Dey, 1994) show that these statistics can be computed by only one fit to the whole dataset. The sum of the n log CPO_i statistics has been termed the LPML. The APT model, with an LPML of -156, does better than the MPT and Dirichlet process models which had LPML values of -162 and -165, respectively; the larger LPML indicates that the APT model predicts the observed data better than the MPT and DPM models. Here and elsewhere we rerun LPML computations multiple times to assure stability.

3.3.2 Ache Armadillo Hunting

Consider data from extended forest treks of the Paraguayan Ache tribe, collected by McMillan (2001) over the course of a year, including armadillo hunting. The data, also considered in Hanson (2006), consists of $n = 38$ adult male hunters' age in years a_i, number of armadillos killed y_i, and number of days spent trekking N_i for hunters $i = 1, 2, \ldots, 38$. The variable of interest is the number of armadillos killed; this quantity contributes to an Ache male's status in the tribe. A quadratic function of the hunter's age a_i plus a hunter-specific random effect x_i is considered as a model for the armadillo kill rate λ_i. The hunter-specific random effects $x_i = \sigma t_{q_i}$ are modeled nonparametrically with $G|c \sim APT_5$ and $c \sim \Gamma(5, 1)$. The model for $i = 1, \ldots, 38$ is $y_i \sim \text{Pois}(N_i \lambda_i)$ with

$$\log(\lambda_i) = \gamma_0 + \gamma_1(a_i - 50) + \gamma_2(a_i - 50)^2 + x_i = \mathbf{z}_i' \boldsymbol{\gamma} + x_i,$$

where $\mathbf{z}_i = (1, a_i - 50, (a_i - 50)^2)'$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)'$. The number 50 is subtracted from each hunter's age to reduce collinearity between the linear and quadratic parts of the transformed kill rate.

The LPML and effective sample size (ESS) (Sargent et al., 2000), are used to compare approaches. The ESS describes the efficiency of mixing during MCMC;

after a suitably long burn in (often 5,000 or more iterates), the ESS measures how many essentially, or "pseudo" *iid* samples are taken from the posterior out of 10,000. The APT GLMM yields a LPML of -98 and ESS values of $1145, 1419, 1953, 2157$ for γ_0, γ_1, γ_2, σ respectively. This model yields posterior means of -0.782, -0.014, and -0.003 for γ_0, γ_1, and γ_2 with posterior standard deviations 0.146, 0.012, and 0.001, respectively. Figure 3.3 shows the empirical kill rates y_i/N_i versus age a_i for the 38 hunters along with the fitted kill rate $\exp\{\hat{\gamma}_0 + \hat{\gamma}_1(a-50) + \hat{\gamma}_2(a-50)^2\}$.

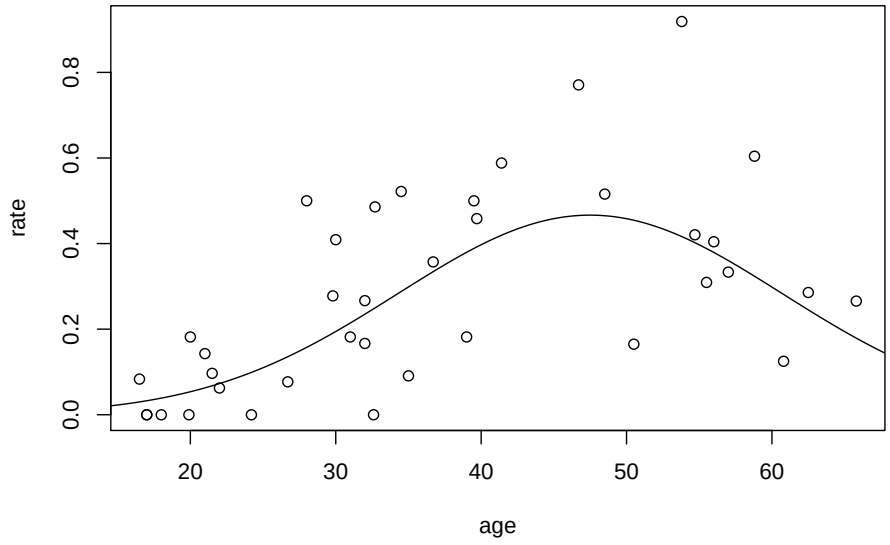

Figure 3.3: Empirical armadillo kill rates y_i/N_i versus age a_i for the 38 Ache hunters. Superimposed is the fitted kill rate from the random intercept APT Poisson model.

Note that the random intercept Poisson model only accommodates extra-Poisson variability; underdispersed data need to be handled differently, e.g. Canale and Dunson (2011). Computing the LPML for the Ache data without random effects yields an LPML of -122, giving evidence of overdispersion and favoring the APT random intercept model. In fact, the pseudo Bayes factor favors the APT model by $\exp(-98+122) > 10^{10}$ compared to the Poisson model without random effects, giving strong evidence of extra-Poisson variability.

We compared the outcome from the APT approach to the results from using a MPT prior described in Hanson (2006). This model can be fit using the `PTglmm` function in the `DPpackage` for `R`. With $c \sim \Gamma(5, 1)$, $J = 5$, and non-informative priors on γ and σ this model yields a LPML of -99 and ESS values of $623, 551, 1021, 675$ for γ_0, γ_1, γ_2, σ respectively. The MPT approach estimates γ_0, γ_1, and γ_2 as -0.724, -0.0179, and -0.003 with standard deviations of 0.133, 0.011, and 0.001, respectively.

Though the coefficient estimates are similar, we see the models are slightly different and the improvements our method provides can be seen in two instances; the LPML statistic shows slightly improved fit of APT relative to MPT while the ESS values show that the mixing is much improved – the APT approach being two to three times greater in terms of efficiency than the MPT approach. A sensitivity analysis on the choice of J reveals there is nothing to be gained from increasing for $G \sim APT_J$. In a similar vein, Agresti (2002) in Section 13.2 notes that often only a handful of discrete mass points are needed for generalized linear mixed models; adding additional discrete mass points leaves the value of the maximized likelihood virtually unchanged after a certain point.

3.3.3 Density estimation

The SAPT is compared to the DPM and the MPT for density estimation. Consider one hundred samples of size $n = 100$ and $n = 500$ from the densities described in Hanson and Jara (2013). The four densities considered are a uniform, Gaussian, double exponential, and a mixture of two Gaussians and a uniform density. For each, the density is estimated using the MPT, DPM and SAPT approaches. The median and 90% interval of L_1 distances from the 100 estimates to the truth are used to compare the three approaches as well as and the median and 90% interval of LPML values. Recall that the L_1 distance is twice the total variation norm, which is interpreted as "how much probability mass needs to be moved from the density

estimate to get the true density." The density estimates averaged over the one hundred replications and true densities are in Figures 3.4 and 3.5.

Both the MPT and SAPT models were fit assuming $c \sim \Gamma(1,1)$ and $J = 6$. The APT further assumes $p(\mu, \sigma) \propto I\{\sigma > 0\}$; the MPT model uses Jeffreys' prior for the underlying normal distribution. The DPM model fit in `DPpackage` is, for $\mathbf{y}_i \in \mathbb{R}^k$,

$$\mathbf{y}_i | G \overset{iid}{\sim} \int N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}) dG(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad G | c, \mathbf{m}_1, \kappa, \boldsymbol{\Psi}_1 \sim DP(c, N_k(\mathbf{m}_1, \tfrac{1}{\kappa}\boldsymbol{\Sigma}) \times IW_k(\nu_1, \boldsymbol{\Psi}_1)),$$

$$\mathbf{m}_1 \sim N_k(\mathbf{m}_2, \mathbf{S}_2), \quad \kappa \sim \Gamma(\tfrac{\tau_1}{2}, \tfrac{\tau_2}{2}), \quad \boldsymbol{\Psi}_1 \sim IW_k(\nu_2, \boldsymbol{\Psi}_2).$$

The inverted-Wishart prior $\mathbf{W} \sim IW_k(\nu, \boldsymbol{\Psi})$ is parameterized such that $E(\mathbf{W}) = \frac{1}{\nu - k - 1}\boldsymbol{\Psi}^{-1}$. The model used here for univariate data is the special case where $k = 1$; the inverted-Wishart reduces to the inverse-gamma in this case. The more general multivariate model is presented above as the data-driven prior presented next works for multivariate density estimation as well.

Let $\bar{\mathbf{y}}$ and $\mathbf{S}$ be the sample mean and covariance matrix for the data $\mathbf{y}_1, \ldots, \mathbf{y}_n$. Let $(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ be a draw under the DPM centering distribution. Then

$$E(\boldsymbol{\Sigma}_j) = E\{E(\boldsymbol{\Sigma}_j | \boldsymbol{\Psi})\} = E\{\boldsymbol{\Psi}^{-1}/(\nu_1 - k - 1)\} = \frac{\nu_2 \boldsymbol{\Psi}_2}{\nu_1 - k - 1}.$$

It makes sense to center the Dirichlet process component covariance matrices on some fraction of $\mathbf{S}$, say $r\mathbf{S}$ to keep the scale and overall shape consistent with the data. For $\boldsymbol{\Sigma}_j$ to have a finite mean, the smallest ν_1 can be is $k + 2$. Thus we take $\nu_2 = k + 2$, $\nu_1 = k + 2$ and $\boldsymbol{\Psi}_2 = r\mathbf{S}$ implying $E(\boldsymbol{\Sigma}_j) = r\mathbf{S}$; we take $r = 0.1$ to have component variances be on the order of one-tenth of the population variances. Now, note that $\boldsymbol{\mu}_j | \mathbf{m}_1, \kappa, \boldsymbol{\Sigma}_j \sim N_k(\mathbf{m}_1, \frac{1}{\kappa}\boldsymbol{\Sigma}_j)$. To allow for cluster means to be located throughout the data cloud, take $\kappa \approx r$, e.g. one could assume $\kappa \sim \Gamma(0.1, 1)$ so $E(\kappa) = 0.1$. Escobar and West (1995) note that $\kappa \to 0^+$ gives bumpier estimates and suggest $\kappa \sim \Gamma(1, 100)$, which we also adopt. Finally, $\mathbf{m}_2 = \bar{\mathbf{y}}$ and $\mathbf{S}_2 = \mathbf{S}$. In this univariate case, we take $v_2 = 3$, $v_1 = 3$, $\Psi_2 = rs^2$ with $r = 0.1$, $\kappa \sim \gamma(1, 100)$ and $m_2 = \bar{y}$ and $s_2 = s^2$, with $\bar{y}$ and s^2 denoting the sample mean and variance of the data, respectively.

In all three models, a burn-in of $1,000$ was used and $3,000$ iterates were saved after burn-in without thinning. The results in Table 3.1 and Figure 3.4 show that all

three models struggle to capture a decent density estimate for the uniform sample with only $n = 100$ observations but generally do well in all other cases. The SAPT generally does about the same or better, relative to the MPT approach, in terms of the L_1 distance across the four densities – the one exception being the spiked double exponential density with sample size of $n = 100$. The SAPT does about the same as the DPM in many cases; overall, though, the DPM provides better L_1 than SAPT or MPT.

Although the SAPT approach does not always yield the best result, we see that it is an improvement upon the MPT approach and provides a real alternative to the DPM model. Hanson and Jara (2013) showed that the MPT approach could be "a serious competitor" to the DPM model; the MPT approach, though, often suffers the disadvantages of yielding quite spiky density estimates. The smoothed APT approximation mitigates that, to an extent, while simultaneously creating a serious competitor to the DPM model.

Table 3.1: Median L_1 and LPML summary over 100 simulations for $n = 100$ and $n = 500$ with a 90% interval.

	DPM		MPT		SAPT	
$n = 100$	LPML	L_1	$LPML$	L_1	$LPML$	L_1
Uniform	-122(-127,-115)	.28(.20,.38)	-123(-128,-117)	.31(.24,.38)	-123(-127,-116)	.30(.24,.38)
Gaussian	-144(-154,-132)	.10(.03,.22)	-144(-155,-132)	.16(.09,.24)	-144(-158,-132)	.16(.09,.24)
Double Exp.	-175(-192,-160)	.20(.12,.31)	-174(-190,-157)	.19(.12,.28)	-175(-190,-159)	.20(.14,.31)
Mixture	-115(-129,-100)	.20(.13,.29)	-121(-135,-103)	.33(.26,.40)	-117(-130,-100)	.23(.17,.31)
$n = 500$	LPML	L_1	$LPML$	L_1	$LPML$	L_1
Uniform	-573(-593,-563)	.16(.12,.21)	-579(-590,-570)	.21(.17,.25)	-576(-587,-567)	.20(.16,.23)
Gaussian	-711(-738,-690)	.04(.02,.09)	-713(-739,-692)	.09(.06,.13)	-713(-739,-689)	.09(.06,.12)
Double Exp.	-850(-891,-821)	.10(.07,.14)	-850(-893,-821)	.13(.09,.17)	-853(-885,-817)	.12(.08,.15)
Mixture	-555(-587,-525)	.10(.06,.14)	-568(-605,-536)	.20(.16,.25)	-564(-591,-529)	.12(.09,.16)

3.3.4 Galaxy Velocities Data

Among many others, Roeder (1990) and Escobar and West (1995) analyzed a dataset on the velocities of eighty-two galaxies sampled from six well-separated conic sections of the Corona Borealis. We estimate the density of these data using the usual MPT, DPM, and SAPT models described in previous sections; a graph of the

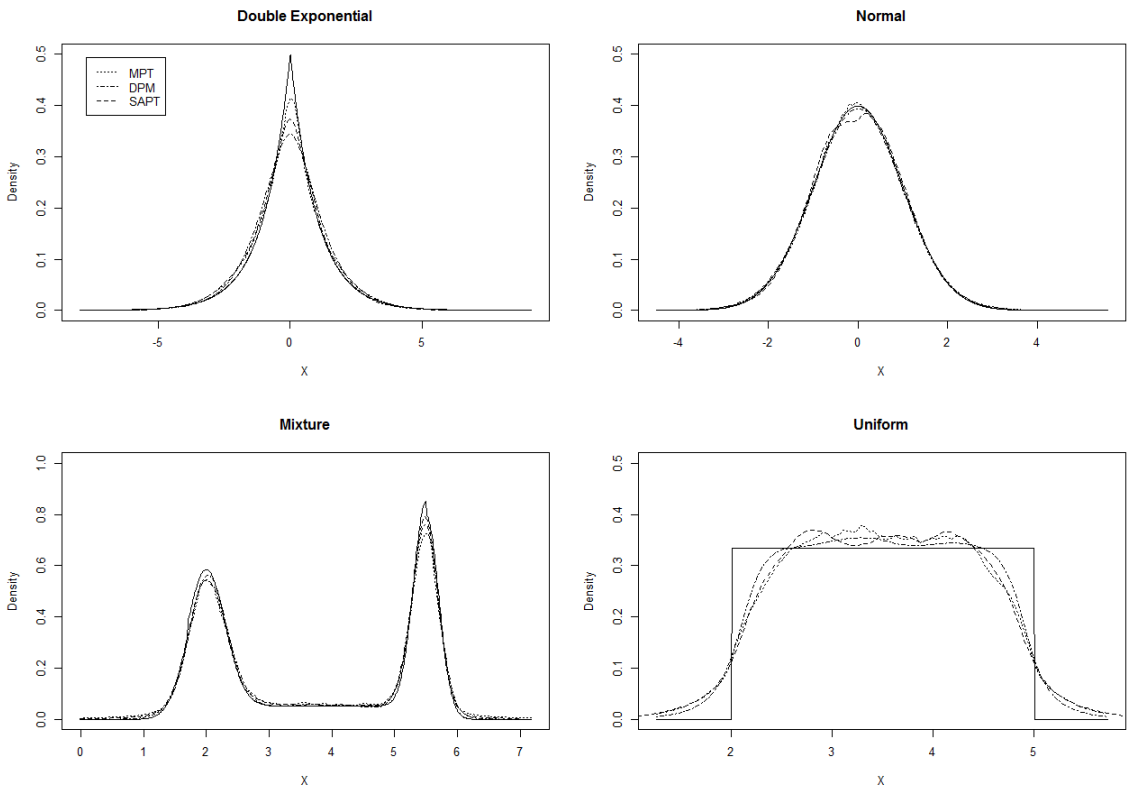

Figure 3.4: Four densities estimated using the MPT, DPM and APT approaches, $n = 100$.

density estimates can be seen in Figure 3.6.

The DPM model was fit using the `DPdensity` function in `DPpackage`. As we did in the previous subsection, we take $v_2 = 3$, $v_1 = 3$, $\Psi_2 = rs^2$ with $r = 0.1$, $\kappa \sim \gamma(1, 100)$ and $m_2 = \bar{y}$ and $s_2 = s^2$, with $\bar{y}$ and s^2 denoting the sample mean and variance of the data.

For the MPT model a burn-in of $10,000$ was used and another $200,000$ iterates were thinned from 20,000,000. For the DPM model a burn-in of $5,000$ was used and another $200,000$ iterates were thinned from 2,000,000. For the SAPT model a burn-in of $5,000$ was used and another $200,000$ iterates were saved without thinning.

The MPT approach yields a LPML of -223 with ESS values of 17 and 11 for μ and σ respectively (out of 10,000). The SAPT approach yields a LPML of -215 with ESS values of 345 and 107 for μ and σ respectively. The DPM approach yields a LPML of

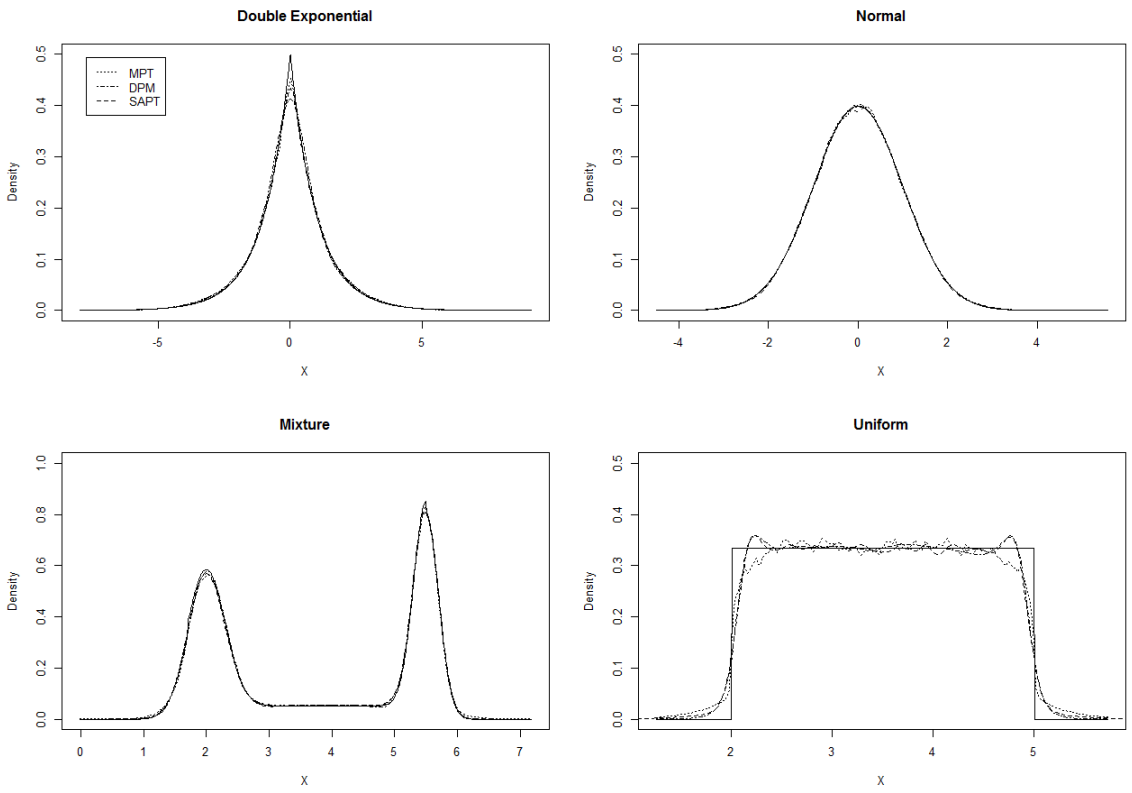

Figure 3.5: Four densities estimated using the MPT, DPM and APT approaches, $n = 500$.

-212 with ESS values of $5,940$ and $5,186$ for the mean and scale, respectively, of the distribution on the Gaussian component means (quite different from the centering distribution parameters of the MPT and SAPT approaches).

The LPML statistic shows that the DPM approach fits the Galaxy data best but the SAPT approach fits similarly and is much improved compared to the MPT approach. The ESS values show that the SAPT mixing is much improved upon the MPT – the APT approach being up to an order of magnitude greater in terms of efficiency than the MPT approach; the DPM has better mixing for the Gaussian component mean centering distribution parameters.

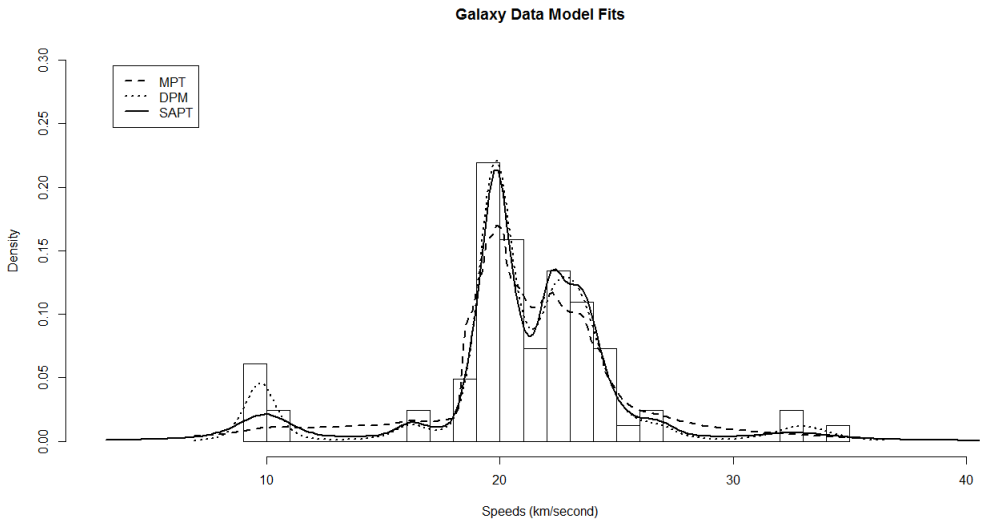

Figure 3.6: Galaxy data histogram and density estimates across models.

3.3.5 Breast cancer survival in Louisiana

The model of Section 3.2.5 is considered for breast cancer survival data from the Surveillance, Epidemiology, and End Results (SEER) Program for the state of Louisiana. The time window considered is 2000-2010, and restricted to those aged 65 years and over at time of diagnosis. Variables of interest include: age at diagnosis, marital status (single, married or other), SEER summary stage (2000 year version; 0=in situ, 1=local, 2=regional, 3=distant), race (white or black) and survival in months after diagnosis (a small number was added to each event time as there were a few zeroes). Observations with missing values for any of these variables are excluded from this study. The dataset is highly right censored, 2212 censored out of 2611 total, i.e. 399 observed deaths. The MCMC scheme of Section 2.5 was implemented in FORTRAN 90 assuming $J = 6$, $c \sim \Gamma(5, 1)$, and $p(\boldsymbol{\gamma}, \sigma) \propto I\{\sigma > 0\}$; 100,000 iterates were thinned to 5,000 after a burn-in of 5,000. The nonparametric approach to censored regression data of Buckley and James (1979) was also applied to these data, using the `bj` function in the `rms` package (Harrell, 2015) for R.

Table 3.2 records the estimated regression effects from both approaches; they

agree fairly well but there are some differences. Consider acceleration factors from the SAPT model in Table 3.2. Whites survive about 1.63 times longer than blacks, holding everything else constant. Going from in situ, to local, to regional, to distance significantly shortens survival as one would expect; significance here meaning that the 95% CI does not contain zero. Age is also significant, with those older at diagnosis tending to live a shorter amount of time. There is no significant effect due to marital status. Survival curves (not shown) show essential agreement for the SAPT, Buckley-James, and censored log-normal models for different covariate combinations, although the parametric log-normal curves are smoother. The LPML for the SAPT model can be computed for censored data using the approach in Ibrahim et al. (2001); the LPML is computed for the log-normal model via the quasi-likelihood method of Geisser and Eddy (1979). These values are -1340 and -1376 respectively on the log-data scale; the SAPT model predicts the data considerably better than the parametric log-normal model.

Table 3.2: SEER Louisiana breast cancer survival in months. Est. is posterior mean, s.d. is posterior standard deviation, and AF is acceleration factor for SAPT model. Est. and s.e. are estimates and standard error from Buckley and James (1979) approach.

	SAPT			Buckley & James	
Effect	Est.	s.d.	AF	Est.	s.e.
Intercept	12.85	0.83		13.24	1.56
Married vs. Single	0.321	0.241	1.38	0.431	0.317
Other vs. Single	-0.00262	0.2209	1.00	0.0670	0.3011
Local vs. In situ	-1.414	0.437	0.243	-1.194	1.201
Regional vs. In situ	-3.035	0.463	0.048	-2.976	1.199
Distant vs. In situ	-5.876	0.496	0.0028	-6.114	1.208
White vs. Black	0.487	0.121	1.63	0.501	0.187
Age	-0.0626	0.0087	0.939	-0.0695	0.0118

3.4 Discussion

A discrete approximation to the univariate MPT model is proposed that has conjugate and efficient MCMC updating for all parameters. A smoothed version, the SAPT, maintains conjugate updating and mixes significantly better than the tradi-

tional MPT. Various simulated and real data illustrate the usefulness of the proposed approach. One obvious extension of the current work is to implement fast estimation of the posterior mode of the SAPT density via the expectation-maximization algorithm (Dempster et al., 1977).

Multivariate versions of the APT and SAPT models suffer the same fate as multivariate extensions of B-splines: an exponential explosion of parameters occurs with increased dimensionality. For example the nonparametric random effects distribution of Ghidey et al. (2004) is largely relegated to bivariate densities. Multivariate Polya trees have largely gotten around this through marginalization; this may be a possibility for the (S)APT as well.

Chapter 4

Supervised Learning using the Polya Trees

The goal in any classification schema is to design a system that classifies new observations into their true class as often as possible. The nonparametric approach of the multivariate Polya tree proposed here has realized impressive results in simulations and real data analyses performing similarly to or better than current approaches in many cases. The flexibility gained from eliminating certain distributional assumptions from the model can greatly improve the ability to correctly classify new observations; even minor deviations from distributional assumptions could lead to missing an important feature in any one class's density. The proposed method is quite fast compared to other supervised classifiers and very simple to implement as there are no kernel tricks or initialization steps that greatly affect the model, like the kernel trick for SVM.

4.1 Introduction

The process of classification involves defining a system of placing a new observation into a level of qualitative variable termed a class. The goal is to choose a system that classifies new observations into their true class as often as possible.

A nonparametric prior that includes the Dirichlet process as a special case is the Polya tree prior (Lavine, 1992). The Bayesian nonparametric Polya tree has realized impressive results in multiple testing (Cipolli et al., 2016), approximate and smoothed density estimation (Cipolli and Hanson, 2016), and density estimation in Euclidean space (Wong and Ma, 2010). The Polya tree prior was initially summarized by Fergu-

son (1974), and further developed by Lavine (1992, 1994), and Mauldin et al. (1992). Hanson (2006) discusses inference for mixtures of finite Polya trees, which smooth out the effect of the partition on posterior inference. In much research regarding Polya trees, however, only univariate data were considered. Paddock et al. (2003) introduced a randomized multivariate Poyla tree which is smoothed over partitions using a random "jitter;" Hanson (2006), rather, introduced a location-scale mixture of Polya trees which directly generalizes the univariate mixture of Polya trees; Hanson et al. (2008) used multivariate Polya trees to model receiver operating characteristic (ROC) curves for evaluating diagnostic test accuracy; Jara et al. (2009) proposed using multivariate Polya trees in generalized linear mixed effect models to remedy the case when the assumption that random effects terms follow a multivariate Gaussian distribution is faulty; Hanson et al. (2011) developed a simple, computationally cheap sampling method for exploring multivariate densities. Müller and Rodriguez (2013) provide a nice summary of various versions of Polya trees.

Specifically, multivariate Polya trees are considered for use in a classification scheme with supervised learning. We marginalize Jara et al. (2009) making computing fast and easy. Consider $y \in Y = \{1, \ldots, g\}$ to be the class of an observation with feature set $\mathbf{x} = \{x_1, \ldots, x_p\} \in \mathbb{R}^p$. Suppose we have data for g classes,

$$\begin{gathered}
\mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1n_1} \\
\mathbf{x}_{21}, \mathbf{x}_{22}, \ldots, \mathbf{x}_{2n_2} \\
\mathbf{x}_{31}, \mathbf{x}_{32}, \ldots, \mathbf{x}_{3n_3} \\
\vdots \\
\mathbf{x}_{g1}, \mathbf{x}_{g2}, \ldots, \mathbf{x}_{gn_g},
\end{gathered}$$

where each $\mathbf{x}_{yi}$ is a $p \times 1$ observation vector of continuous observations corresponding to class $y \in Y$. The goal, then, is to classify a new $p \times 1$ observation vector, $\mathbf{x}_0$, as a member of one of the g classes.

In section 2, we provide a literature review of many popular supervised learning classification systems. Many of these classification approaches make distributional assumptions on the observations $\mathbf{x}_{yi}$, perhaps most infamously assuming each $\mathbf{x}_{yi}$ is drawn from a multivariate Gaussian distribution, i.e. $\mathbf{x}_{yi} \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$. A more flexible approach can be obtained by changing the Gaussian, or any distributional assumption for that matter, to the nonparametric multivariate Polya tree (Hanson, 2006). More specifically, observations $\mathbf{x}_{yi}$ follow a multivariate Polya Tree distribution centered at a multivariate Gaussian with mean $\boldsymbol{\mu}_y$ and covariance matrix $\boldsymbol{\Sigma}_y$.

The flexibility gained from eliminating the distributional assumptions from our analysis can greatly improve our ability to correctly classify a new observation; even minor deviations from the distributional assumptions could lead to missing an important feature in any one classes' density. This proposed specification would not only remedy the case where the all the classes' distributions are not Gaussian but also allows each level to have vastly different densities from each other. The ability of the Polya tree approach to pick out even slight deviations significantly improves classification over a model that is assumption heavy particularly in the case when the data significantly deviates from those assumptions.

We introduce a supervised learning technique utilizing the Polya tree distribution. We fit the data with a multivariate Polya tree prior to each class y and use the estimated multivariate density, $f_y(\mathbf{x})$, to classify a new feature vector $\mathbf{x}_0$ to the class denoted as most likely by the conditional distribution of $\mathbf{x}$ given the training data from each class through Bayes' rule. The outcome for analysis with this model is compared to other methods including nearest neighbor, naive Bayes, artificial neural networks, and support vector machines.

4.2 Literature Review

There are many literature reviews and textbooks available on the topic of statistical learning i.e. Duda et al. (2000) Hastie et al. (2001), Larrañaga et al. (2006), Kotsiantis (2007), Mohri et al. (2012), and Alpaydin (2014). Below we briefly summarize techniques similar the one proposed herein and those classic approaches that we compare results among.

4.2.1 Bayes' Classifier

Often used as a benchmark for comparison in simulation studies where the truth is known, the Bayes' classifier minimizes test error on average. This classifier assigns each new observation to the most likely class via Bayes' rule; for example, Bayes' classifier assigns a new $p \times 1$ observation vector $\mathbf{x}_0$ into class

$$arg\max_y p(Y = y|\mathbf{X} = \mathbf{x}_0) \propto \arg\max_y p(Y = y)p(\mathbf{x}_0|Y = y) = \arg\max_y \pi_y f_y(\mathbf{x}_0),$$

where $f_y(\cdot)$ is the distribution of $\mathbf{x}$ given class $y \in Y$ and π_y is the population proportion of observations in class $y \in Y$. Though this classifier produces the lowest average test error rate, it is impossible to use in practice since it is required to know from what distribution the observations $\mathbf{x}$ are drawn across each class so that the conditional probability can be evaluated for each new observation $\mathbf{x}_0$ as well as the population proportion.

The test error for classifying $\mathbf{X} = \mathbf{x}_0$ is the complement of the original rule, i.e. $1 - \arg\max_y \pi_y f_y(\mathbf{x}_0)$ and so the expected test error is

$$1 - E\{\arg\max_y \pi_y f_y(\mathbf{x}_0)\}.$$

It should be noted that this error is usually not zero as the populations typically overlap in which case this can be considered irreducible error; hence the Bayes' classifier is used as a benchmark.

4.2.2 Kernel Density Estimation Classifiers

Kernel density estimation classifiers use kernel density estimates of the underlying probability density function to classify new observations $\mathbf{x}_0$ into the most probable class Y. These estimates are then used as Bayes' classifiers when it is the case that the densities are unknown. There are many approaches but perhaps the most simple and understandable case would be to use a histogram to estimate a univariate probability distribution function. In this case we must choose the support to consider and the bin size, noting that this will lead to a piecewise linear density estimate.

The smoothing process requires an appropriate choice of Kernel. There are many choices of kernel, i.e. Epanechnikov, Tri-cube, Gaussian etc., to smooth out the points to accommodate various requirements on support, continuity, and differentiability. Perhaps more important than the Kernel selection is the bandwidth selection which has a variance-bias tradeoff: a larger bandwidth lowers variance with higher bias whereas a lower bandwidth increases variance with lower bias.

A review of techniques of this type can be found in Izenman (1991), Ledl (2004), Marzio and Taylor (2005), Mukhopadhyay and Ghosh (2011), Zambom and Dias (2013), and Chapter six of Hastie et al. (2001).

4.2.3 Naive Bayes' Classifier (NB)

The naive Bayes' classifier model is also based on Bayes' theorem and is very closely related to the benchmark Bayes' classifier; the "naive" addition refers to the added assumptions of conditional independence and not knowing the densities for each class y. In order to calculate the class probabilities outlined above, when talking about the Bayes' Classifier, we require a prior and likelihood.

In the case of the Bayes' Classifier we classify new observation $\mathbf{x}_0$ into group i that maximizes the conditional probability from Bayes' Rule. The naive Bayes' classifier is developed via the following use of the "naive" independence assumption, $f_y(\mathbf{x}_0)$ is

calculated and new observation $\mathbf{x}_0$ is classified into class

$$\arg\max_y \pi_y f_y(\mathbf{x}_0) = \arg\max_y \pi_y \prod_{k=1}^{p} f_{yk}(x_{0k}),$$

where $f_{yk}(x_{0k})$ is the marginal density of x_{0k}, the k^{th} feature, under group y.

The population proportion of observations in group $y = 1, \ldots, g$ is estimated by the sample proportions in the training data unless they are known to be otherwise and the densities are estimated in a variety of ways, the choice of which largely depends on the type of data being analyzed. For example, when dealing with continuous data it is often assumed that each feature is drawn from a Gaussian distribution parameterized by mean and variance estimated by the sample mean and variance calculated from the training data; i.e. $f_{yk}(x_{0k}) = N(\overline{x_{yk}}, s^2_{yk})$ where $\overline{x_{yk}}$ is the sample mean and s^2_{yk} is the sample variance of the k^{th} feature in group y as calculated from the training data.

Rish (2001) provides a nice empirical study of the Naive Bayes' Classifier and Jiang et al. (2007) provides a survey of improved versions of the Naive Bayes' Classifier. Daumé III and Marcu (2005) introduce a Bayesian approach based on the Dirichlet process prior, which is implemented via Markov chain Monte Carlo techniques extending the standard naive Bayes' classification model to the case where the number of classes is allowed to grow unboundedly.

4.2.4 Gaussian Process Classification

Gaussian process priors are widely used models across different fields, including as a binary classification method. The classification model works by defining latent variables for feature sets $\mathbf{x}$ and modeling the posterior probabilities; in terms of the Bayes' Classifier, the Gaussian process classifier computes the class probabilities by considering the probability of being in class y conditional on the defined latent variables.

A Gaussian process defines a distribution over functions; in terms of the Bayes' classifier we let $\gamma = \{\gamma_1, \ldots, \gamma_p\}$ be random, latent function variables that correspond to the set of inputs $\mathbf{x} = (x_1, \ldots, x_p)'$ where $p(\gamma|\mathbf{x})$ has a multivariate Gaussian distribution. For classification, place a Gaussian process prior on $f_y(\mathbf{x}_0)$, as above, and use a likelihood, e.g. a sigmoid function, to consider $p(Y = y|\mathbf{X} = \mathbf{x}_0)$. Since the choice of likelihood is not Gaussian, integrating over γ to evaluate $f_y(\mathbf{x}_0)$ is intractable creating the need for approximating methods like Markov chain Monte Carlo (Atiya et al., 2013), expectation propagation (Minka, 2001) or Laplace approximation (Williams and Barber, 1998); several other approximation and sampling based methods are summarized in Nickisch and Rasmussen (2008). Regardless of the method used for approximation, new observation $\mathbf{x}_0$ is classified into class $\arg\max_y \pi_y f_y(\mathbf{x}_0)$.

Rasmussen and Williams (2006) provides a review of theory and application of Gaussian processes in supervised-learning for both regression and classification from both a Bayesian and frequentist point of view. This text reviews several approximation methods mentioned above as well as the issue of having to choose a covariance, or kernel, function. It is also noted that Gaussian process classification has many connections to other well-known learning techniques, including support-vector machines, neural networks, etc.

4.2.5 Dirichlet Process and Dirichlet Process Mixture Classification Models

Shahbaba and Neal (2009) introduced a nonlinear model for classification over two or more classes, as well as over classes that have hierarchical structure. This approach uses Dirichlet process mixtures with centering distribution G_0 to nonparametrically estimate the joint distribution of $(Y, \mathbf{X})$ under the assumption that the covariates of $\mathbf{X}$ are independent and that the dependence between Y and $\mathbf{X}$ can be modeled using a linear model within each class.

They use a Gaussian mixture to model the covariates of $\mathbf{X}$ and a $g-$class multinomial logistic model for the class variable Y. Each class in the mixture model has

parameters $\theta = \{\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$ and the distribution of $\mathbf{X}$ within each class is multivariate normal with mean and diagonal covariance, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ respectively. The full model is: $P \sim DP(\alpha G_0)$; $\theta_y|P \sim P$; $x_{yi}|\theta_y \sim N(\mu_{yi}, \sigma^2_{yi})$. Finally, for a new observation $\mathbf{x}_0$ is classified into class

$$\arg\max_y p(Y = y|\mathbf{x}_0, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{exp(\alpha_y + \mathbf{x}_0'\boldsymbol{\beta}_y)}{\sum_{k=1}^{g} exp(\alpha_k + \mathbf{x}_0'\boldsymbol{\beta}_k)};$$

Note that the parameters α_k are scalars and $\boldsymbol{\beta}_k$ are $p \times 1$ vectors of the GLM parameters.

Hannah et al. (2011) generalizes this by proposing a Dirichlet process mixtures of GLMs, based on the Chinese Restaurant Process, which produces an asymptotically unbiased estimate of the mean function given input-response pairs, $(Y, \mathbf{X})$. That same mean function is then used to map each new observation $\mathbf{x}_0$ to an average response, or class. This approach generalizes the approach of Shahbaba and Neal (2009) to a wider selection of response distributions.

4.2.6 Linear Discriminant Analysis (LDA)

LDA for discriminating between two classes was introduced by Fisher (1936) and extended to multiple classes by Rao (1948). LDA constructs a classifier based on a linear combination of Gaussian distributed features that best separates two classes. In this approach it is assumed that the observations across classes are independently Gaussian distributed with different means, $\boldsymbol{\mu}_y$, but a common covariance matrix, $\boldsymbol{\Sigma}$. Observations are classified by projecting data from a p-dimensional feature vector onto an the line which best separates the samples from a probabilistic sense.

The idea is mathematically simple, computationally efficient and still beats some approaches. LDA, however, is quite inflexible as it makes many model assumptions and though it works well when all the assumptions are met it typically underperforms compared to other classification techniques.

LDA is derived via Bayes' rule:

$$p(Y = y|\mathbf{X} = \mathbf{x}) \propto f_y(\mathbf{x})\pi_y = \frac{1}{(2\pi)^n\sqrt{|\boldsymbol{\Sigma}|}} e^{\left(.5(\mathbf{x}-\boldsymbol{\mu}_y)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_y)\right)}$$

where the population proportion of observations in group $y = 1, \ldots, g$ and Gaussian parameters are estimated by the sample values in the training data. This reduces to an LDA classification that assigns new observation $\mathbf{x}_0$ to class

$$\arg\max_y \boldsymbol{\mu}_y \boldsymbol{\Sigma}^{-1} \mathbf{x}_0' - .5\boldsymbol{\mu}_y \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_y' + ln(\pi_y).$$

4.2.7 Quadratic Discriminant Analysis (QDA)

QDA provides an extension to LDA by still assuming classes are independently Gaussian distributed with different means, $\boldsymbol{\mu}_y$, but relaxing the covariance assumption to allow different covariance matrices, $\boldsymbol{\Sigma}_y$ across groups. This classifier is based on a quadratic combination of features thus extending LDA by allowing the classifier to represent non-linear separations.

QDA shares the same benefits as LDA as it is computationally efficient and still beats some algorithms when all the assumptions are met but some of the drawbacks of LDA are mitigated because QDA assumes a more flexible model. However, as the number of predictors increases the computational expense of QDA increases; while using LDA is less computationally expensive it can suffer bias if used when the common covariance assumption is faulty. QDA works quite well when the sample size is large, but if the training set is small LDA tends to work better as the added assumption obtains estimates with lower variance. While QDA is more flexible than LDA it still makes many model assumptions compared to other models and though it works well when all the assumptions are met it typically underperforms compared to other, more advanced classification techniques.

Similar to LDA, QDA is derived via Bayes' rule:

$$p(Y = i|\mathbf{x}) \propto f_y(\mathbf{x})\pi_y = \frac{1}{(2\pi)^n \sqrt{|\boldsymbol{\Sigma}_y|}} exp\left(.5(\mathbf{x} - \boldsymbol{\mu}_y)' \boldsymbol{\Sigma}_y^{-1} (\mathbf{x} - \boldsymbol{\mu}_y)\right)$$

where the population proportion of observations in group $y = 1, \ldots, g$ and Gaussian parameters are estimated by the sample values from the training data. This reduces to a QDA classification that assigns new observation $\mathbf{x}_0$ to class

$$\arg\max_y log(|\boldsymbol{\Sigma}_y^{-1}|) + (\mathbf{x}_0 - \boldsymbol{\mu}_y)' \boldsymbol{\Sigma}_y^{-1} (\mathbf{x}_0 - \boldsymbol{\mu}_y) - 2log(\pi_y).$$

4.2.8 k Nearest Neighbor (KNN)

The k nearest neighbor approach, introduced by Cover and Hart (1967), is a nonparametric method for classification. Each observation is classified by considering the k "nearest" points, as evaluated by some distance measure, and assigning the most common class among these neighboring points to the new observation. This approach uses a very simple, but crude method of estimating the conditional probability as described in the Bayes' classifier section by using a popular vote of the nearest points. In terms of the Bayes' classifier, k nearest neighbor assigns new observation $\mathbf{x}_0$ to class $\arg\max_y k^{-1} \sum_{i=1}^{k} I(y_{s_i} = y)$ where $s_1, \ldots, s_n$ are a permutation of $\{1, \ldots, n\}$ that enforce the order $||\mathbf{x}_{s_1} - \mathbf{x}_0|| \leq ||\mathbf{x}_{s_2} - \mathbf{x}_0|| \leq \cdots \leq ||\mathbf{x}_{s_n} - \mathbf{x}_0||$ and $I(Y_{s_i} = y)$ is the indicator function which returns one if observation $\mathbf{x}_{s_i}$ is from class y and zero otherwise. By 'nearest' we mean the first k points where $||\mathbf{x}_i - \mathbf{x}_0||$ is as small as possible and $||\cdot||$ is Euclidean distance.

The model is flexible in that the user can choose the number k and the distance measurement. Euclidean distance is a popular choice but the user is free to define any meaningful distance according to the problem at hand. The choice of k is slightly more complicated. If k is too small the model is overly flexible creating many boundaries and if k is too large the boundary approaches a linear boundary. Solving this problem is not very simple, but there are a couple approaches that lead us to reasonable choices of k, in what follows take the empirical rule-of-thumb $k = \sqrt{n}$ as in Duda et al. (2000). There are many approaches to choosing k including simply choosing k which minimizes cross validation test error or more complicated approaches such as the Bayesian approach of Ghosh (2006) or using ensemble learning (Hassanat et al., 2014).

The model is attractive because it is easy to understand, the implementation is very simple, there is no training and it still achieves good results in basic recognition problems. The issue is that it is quite sensitive to local structure in the data so it

struggles with noise. It is also quite computationally expensive particularly as sample size grows because the distance between every two points must be calculated. Dudani (1976) introduced a distance weighted nearest neighbor approach where neighbors are weighted so that nearer observations bare more weight on the classification.

4.2.9 Random Forest (RF)

The random forest model is an ensemble method that combines predictions made by multiple decision trees. The idea was first introduced by Ho (1995) and then, a later version, by Breiman (2001) which is the model we use in Section 4. The random forests model is a way of averaging multiple decision trees, built randomly by considering a random sample of predictors at each split in the tree instead of the full set. By introducing this methodology the random trees are far less correlated than the trees produced by bagging (Breiman, 1996), which leads to a reduction of the variance, compared to binary trees, at the expense of bias and interpretability. The training is done in a divide and conquer fashion and to generate a prediction a vote among all the underlying trees takes place and the majority prediction value wins.

In terms of classification, the random forest model assigns new observation $\mathbf{x}_0$ to class $\arg\max_y t^{-1} \sum_{j=1}^{t} I(f_j(\mathbf{x}_0) = y)$ where $f_j(\mathbf{x}_0)$ returns the classification from decision tree j and $I(\cdot)$ is an indicator function which returns 1 if decision tree j determines $\mathbf{x}_0$ is from class y and 0 otherwise.

The random forest model has become popular because deep trees can learn highly irregular patterns, even with much noise, quite quickly. This approach is also able to deal with unbalanced or missing data and is does surprisingly well with small sample sizes. Random forests cannot, however, predict for features beyond the range in the training data which affects generalization and are difficult to interpret.

4.2.10 Artificial Neural Networks (ANN)

During the last couple decades ANN has been extensively researched. Below we provide some history and the general idea but many literature reviews, i.e. Anderson and Rosenfeld (1988), and textbooks, i.e. Yegnanarayana (2004) and Rojas (1996), thoroughly explore the theory and application of ANN.

ANN is inspired by attempts to simulate biological neural systems by a connected system. The first artificial neuron was produced by Mcculloch and Pitts (1943) but the technology available at that time hampered further research and application. This approach was limited to modeling logical expressions, like the mathematical "or," for binary inputs and output. Later, the first successful applications were in the form of neuro-computers, i.e. the Mark I perceptron (Rosenblatt, 1958), ADALINE (Widrow and Hoff, 1960) and MADALINE (Widrow, 1962). After a period of inactivity, largely due to the critique of Minsky and Papert (1969) who pointed out the methods inability to learn the "exclusive or" logical expression, Werbos (1981) founded back propagation which reinvigorated and propelled ANN research.

Current ANN methods provide a flexible, nonparametric approach that do not require any assumptions on the model's structure before application and are capable of modeling highly nonlinear systems. The flexibility of this model allows users to model practically any dataset including applications with multiple simultaneous outputs. These attributes make ANN very popular for applications research and in industry as ANN can be used to recognize patterns and that are too complex and noisy to be realized by other approaches. ANN generally does quite well in discovering patterns, however there are cumbersome tasks to ensure it does well; one needs to select an activation function, the number of hidden layers, the number of nodes in each layer and the weights on the edges that connect them. As the training sample size and the dimensions increase an ANN can quickly become quite computationally intractable - sometimes taking weeks to finish training.

The structure of a basic ANN is represented by a connected graph and consists of three layers - an input layer, a hidden layer and an output layer. The input layer consists of p nodes, one for each dimension of the feature set, and the output layer has g nodes, one for each class. Karsoliya (2012) discusses the selection of the number of nodes in the hidden layer but a rule of thumb is to take $2d/3$ hidden nodes. For calculation feasibility most applications use one hidden layer but increased computation ability has led to deep learning algorithms which use many hidden layers; Schmidhuber (2015) provide a nice summary of more advanced models such as the multilayer ANN, recurrent ANN and deep learning which allow for more complex relationships between the input and output layers making this contribution to research quite important.

In terms of the classification, the ANN model assigns new observation $\mathbf{x}_0$ to class $\arg\max_y f_y(\mathbf{x}_0)$ where $f_y(\mathbf{x})$ is the fitted value of $\mathbf{x}_0$ or output for group y; i.e. $\mathbf{x}_0$ is classified into the group associated with the largest valued output node.

While many consider ANN to be one of the most flexible models, it creates a sort of hazard; ANN can be applied to any problem including those which we may not know very much about. The results from ANN must be very carefully interpreted and compared to expert knowledge because it is a black box learner, meaning one cannot interpret the relationship between any input and output through the model. This is particularly important because ANN can suffer from overfitting and can only guarantee local optima.

4.2.11 Support Vector Machines (SVM)

During the last few decades SVM has been extensively researched. Below we provide some history and the general idea but many literature reviews, i.e. Burges (1998), and text books, i.e. Steinwart and Christmann (2008) and Ma and Guo (2014), thoroughly explore the theory and application of SVM.

The generalized portrait algorithm was introduced by Vapnik and Lerner (1962) and Vapnik and Chervonenkis (1963); this idea was extended to linear SVM by Vapnik (1979). Boser et al. (1992) generalize SVM further to the case where the separation is non-linear by using a kernel trick where it is argued that the kernels can be interpreted as inner products of an expanded feature space, geometrically; Scholkopf and Smola (2001) provide a nice summary of SVMs and associated kernel methods.

SVM searches for and constructs hyperplanes which separate the training data by as much as possible and then uses the hyperplanes as a boundary for classification. This is delightfully simple when datasets are linearly separable and in the case where the separation is nonlinear one can utilize the kernel trick to map the feature set to some higher-dimensional feature space where the training set is separable. SVM is widely praised as the best "out of the box" classifier as it generally classifies very well. It should be noted that as the training sample size and the dimensions increase a nonlinear SVM can quickly become computationally intractable and that this computational cost continues during the test stage. Another drawback is for data that is not linearly separable as it's necessary to select a sophisticated kernel function to employ and to determine reasonable tuning parameters; incorrect choices of kernel can lead to significantly higher classification error as seen in the simulations and data analysis in Section 4.

When data is linearly separable hard-margin (Boser et al., 1992) is used for classification. Geometrically hard-margin is designed to find two parallel hyperplanes that separate the data such that the distance between the two hyperplanes, δ, is maximized. Consider the case where we have two classes $y_1 = -1$ and $y_2 = 1$. We can compute two parallel hyperplanes $(\mathbf{w}'\mathbf{x}+b) = 1$ and $(\mathbf{w}'\mathbf{x}+b) = -1$ with the distance between them $\delta = 2/||\mathbf{w}||$. The hard-margin problem reduces to finding $\mathbf{w}$ and b that minimizes $||\mathbf{w}||$ constricted to the data being linearly separable, i.e. $y(\mathbf{w}'\mathbf{x} + b) \geq 1$. Finally, $\mathbf{x}_0$ is classified into y_1 if $(\mathbf{w}'\mathbf{x}_0 + b) < 0$ and y_2 if $(\mathbf{w}'\mathbf{x}_0 + b) > 0$.

Alternatively, when the separation is non-linear, soft-margin (Cortes and Vapnik, 1995) is used for classification. Similarly we start with two parallel hyperplanes $(\mathbf{w}'\mathbf{x}+b) = 1$ and $(\mathbf{w}'\mathbf{x}+b) = -1$ with slack variables $\epsilon_i \geq 0$ that represent the distance from a misclassified observation to the hyperplanes separating its true class. The soft margin problem reduces to finding $\mathbf{w}$ and b that minimizes $||\mathbf{w}|| + C\sum_{k=1}^{R}\epsilon_k$ where C is a parameter that controls overfitting and there are R misclassified observations, i.e. $y(\mathbf{w}'\mathbf{x}+b) \geq 1-\epsilon$. Note that for small values of C soft-margin acts like hard-margin and large values of C can cause overfitting. Finally, $\mathbf{x}_0$ is classified into y_1 if $(\mathbf{w}'\mathbf{x}_0+b) < 0$ and y_2 if $(\mathbf{w}'\mathbf{x}_0+b) > 0$. On using a kernel function κ this decision rule is altered to classify $\mathbf{x}_0$ into y_1 if $(\mathbf{w}'\kappa(\mathbf{x}_0)+b) < 0$ and y_2 if $(\mathbf{w}'\kappa(\mathbf{x}_0)+b) > 0$.

There are many papers that explore the extension to multiple classes. Hastie and Tibshirani (1998) suggest pairwise coupling and a summary of other methods is given by Duan and Keerthi (2005).

4.3 Models

Hanson (2006) and Hanson et al. (2008) discuss multivariate mixtures of finite Polya trees and a p-dimensional location-scale mixture that is a direct generalization of the univariate finite location-scale mixture. For each level j of the Polya tree we partition $\mathbb{R}^d$ into 2^{jd} partitions; denote the partition at level j as Π_θ^j. Consider data uniformly distributed on the unit square centered at the origin; with $p = 2$ take quaternary splits to make 4^j partitions for each level $j = 1, 2, 3$. The first split is the usual x and y axes of the Cartesian coordinate system making four partitions on level $j = 1$, sixteen on level $j = 2$, sixty four on level $j = 3$. Figure (4.1) shows this partitioning using decreasing line-widths for each level j so one can see the window-pane-effect.

These partitions sets within Π_θ^j are given by:

$$B(j;\mathbf{k}) = (\Phi^{-1}((k_1-1)/2^j), \Phi^{-1}((k_1)/2^j)) \times \ldots \times (\Phi^{-1}((k_d-1)/2^j), \Phi^{-1}((k_d)/2^j))$$

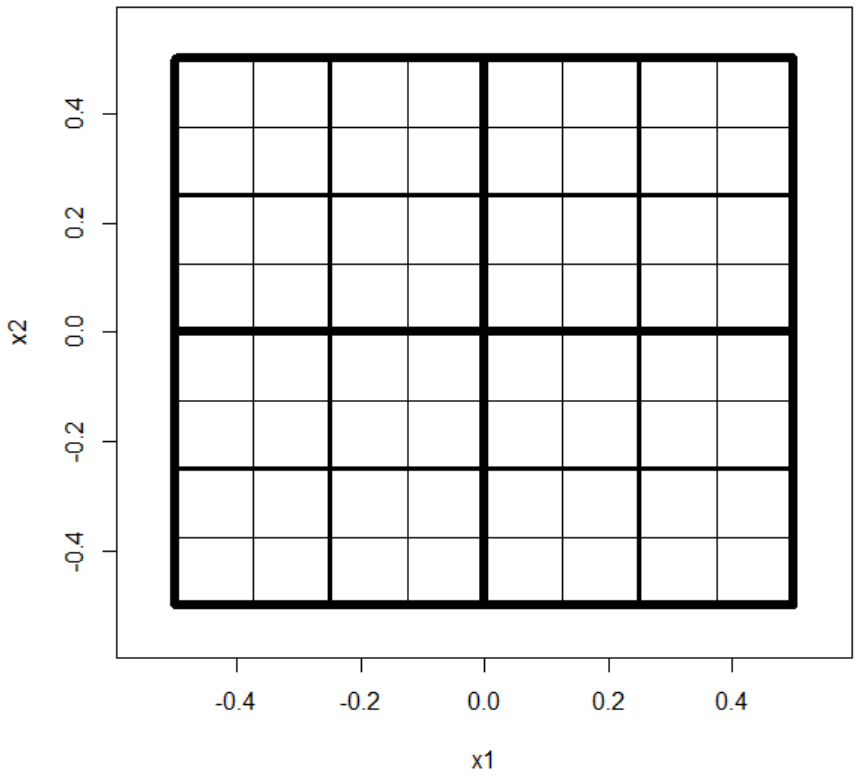

Figure 4.1: Polya tree partitions in $\mathbb{R}^2$ for data uniformly distributed on the unit square centered at the origin

for vectors $\mathbf{k} = [k_1, k_2, \ldots, k_d]$ such that $k_t \epsilon [1, 2, \ldots, 2^j]$ for $t = 1, \ldots, d$; where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function for a standard normal. Note that the 2^{jd} intervals $B(j; \mathbf{k})$ partition $\mathbb{R}^d$ up to a set of Lebesgue measure zero.

A Polya tree is centered at the multivariate Gaussian with mean vector $\boldsymbol{\mu}_y$ and covariance matrix $\boldsymbol{\Sigma}_y$ which are estimated via empirical Bayes' estimates for each group y. We calculate which partition set, $B(j; \cdot)$, each observation, $\mathbf{x}_i$ lies on at each level j using the following $r_{ij} = \sum_{p=1}^{d} 2^{j(p-1)} \lfloor 2^j \Phi(z_{ip}) \rfloor$, where r is a $n_y \times J + 1$ matrix; $\mathbf{z}_i = \boldsymbol{\Sigma}_y^{-1/2}(\mathbf{x}_i - \boldsymbol{\mu}_y)$ is the centered and scaled version of the i^{th} feature set; z_{ip} is the p^{th} dimension of $\mathbf{z}_i$; Φ is the standard Gaussian CDF; $\lfloor . \rfloor$ is the usual floor function; and r_{ij} gives the numbered partition that $\mathbf{x}_i$ lies on. For $j = 0$ set $r_{i0} = 0$.

For a new observation $\mathbf{x}_0$ the conditional probability on each group $y = 1, \ldots, g$

is

$$p_y(\mathbf{x}_0|x_{y1},\ldots,x_{yn_y},\boldsymbol{\mu}_{yk},\boldsymbol{\Sigma}_{yk},c_y)=\phi_d(\mathbf{x}_0)\prod_{j=1}^{J}\frac{2^d c_y j^2+2^d n(j,k(j,\mathbf{x}_0)|x_{y1},\ldots,x_{yn_y})}{2^d c_y j^2+n(j-1,k(j-1,\mathbf{x}_0)|x_{y1},\ldots,x_{yn_y})},$$

where c_y is the parameter that controls how "close" the Polya tree is to its centering distribution, i.e. the multivariate Gaussian distribution, and J is the number of levels of the finite tree. Hanson (2006) suggests taking $J=\lceil log_{2^d}(n_y)\rceil$ for each group. To choose c_y consider the predictive density of i^{th} observation of group y, $\mathbf{x}_{yi}$, given all the observations up to that point $\mathbf{x}_{y\{1:i-1\}}=\mathbf{x}_{y1},\ldots,\mathbf{x}_{y(i-1)}$, $\boldsymbol{\mu}_y,\boldsymbol{\Sigma}_y$ and c_y. This is denoted

$$p(\mathbf{x}_{yi}|\mathbf{x}_{y\{1:i-1\}},c_y,\boldsymbol{\mu}_y,\boldsymbol{\Sigma}_y)=\prod_{j=1}^{J}\frac{2^d j^2 c_y+2^d n(j,k(j,\mathbf{x}_i,\mathbf{x}_{y\{1:i-1\}})}{2^d j^2 c_y+2^d n(j-1,k(j-1,\mathbf{x},\mathbf{x}_{y\{1:i-1\}})}\phi_d(\mathbf{x}|\boldsymbol{\mu}_y,\boldsymbol{\Sigma}_y).$$

To choose c_y, take

$$c_y=\arg\max_{c_y>0}\prod_{i=1}^{n_y}p(\mathbf{x}_i|\mathbf{x}_{y\{1:i-1\}},c_y,\boldsymbol{\mu}_y,\boldsymbol{\Sigma}_y)=\arg\max_{c_y>0}p(\mathbf{x}_{y1},\ldots,\mathbf{x}_{yn_y}|c_y,\boldsymbol{\mu}_y,\boldsymbol{\Sigma}_y),$$

computationally it is quick and simple to check and choose the maximum over the interval from .01 to 100 by .01 increments.

The multivariate density estimate provided by the Polya tree is a bit blocky - extending the concern of spikiness in univariate density estimation of the Polya tree to the p-dimensional case. In the two-dimensional case the bivariate density estimate is boxy where the bivariate heat map of the density suffers from a "window pane effect" where the expected smoothness is lost; this is not dissimilar to the smoothness lost using a bivariate histogram. There are two routes for smoothing by averaging density estimates across rotations. Givens rotations (Golub and Van Loan, 1996) are useful in high dimensions but require MCMC and extensive computation. Another option is to simply consider only a handful of square roots, maybe just two, e.g. $\boldsymbol{\Sigma}_{y1}^{-1/2}=\mathbf{V}\boldsymbol{\lambda}\mathbf{V}'$ and $\boldsymbol{\Sigma}_{y1}^{-1/2}=\mathbf{V}\boldsymbol{\lambda}$, where $\mathbf{V}$ is the matrix of eigenvectors and $\boldsymbol{\lambda}$ is the diagonal matrix of eigen values $\lambda_1,\ldots,\lambda_d$ of the sample covariance as calculated from the training data. We opt for the latter since it is computationally less expensive and can, in many cases, do about the same as the former.

In terms of the Bayes' classifier, classification is completed by assigning $\mathbf{x}_0$ to group

$$\arg\max_y\sum_{k=1}^{2}\pi_y p_{yk}(\mathbf{x}_0|x_{y1},\ldots,x_{yn_y},\boldsymbol{\mu}_{yk},\boldsymbol{\Sigma}_{yk},c_y),$$

where $p_{yk}(\cdot)$ is the conditional probability on each group $y = 1, \ldots, g$ for each square root $\mathbf{\Sigma}_{yk}$, $k = 1, 2$.

4.4 Illustrations

Below "Polya" refers to the proposed methodology, "SVMlin" is the SVM methodology with a linear kernel, "SVMrad" is the SVM methodology with a radial basis function kernel,"SVMpoly" is the SVM methodology with a polynomial kernel, and "SVMsig" is the SVM methodology with a sigmoid kernel; all SVM versions are fit using the `e1071` package (Meyer et al., 2015) for the programming language R (R Core Team, 2014). KNN is fit using the `FNN` package (Beygelzimer et al., 2013), LDA and QDA are fit using the `MASS` package (Venables and Ripley, 2002), ANN is fit using the `nnet`, (Venables and Ripley, 2002) and RF is fit using the `randomForest` package (Liaw and Wiener, 2002) for R.

4.4.1 Red or Blue States

Florida (2011) explores associations to assess how income, education and other factors influence the propensity for international travel. To complete this task median income, the percent of a population with a bachelor's degree and several other variables for each state are considered and the correlation between them and passport ownership is used to argue whether or not there are any associations.

A classification-minded extension to the analysis of Florida (2011) is to consider classifying states as Democrat or Republican based off similar data. There is prominent coverage of this cycle's primary election in the United States at the time of writing making this a timely and interesting example! In the following analysis, the variables that make up a five-dimensional feature set are the percentage of the population who have received a passport between 2009 and 2015 (United States Department of State Bureau of Consular Affairs, 2015); median income, using a three year aver-

age, from 2011 to 2013, in 2013 inflation adjusted US Dollars (United States Census Bureau, 2014); the percent of the population with a Bachelor's Degree or higher in 2010 (United States Census Bureau, 2010); federal funding per US dollar of taxes paid in 2005 (Tax Foundation, 2007); and the percent of foreign born residents (Migration Policy Institute, 2014); for each state. The classes, Democrat or Republican, are decided by how each state voted in the last five presidential elections (National Archives and Records Administration, 2012), i.e. if a state voted Republican three out of five times that state is classified as Republican.

Table 4.1: Numerical summaries for the feature set.

Group	% Passport	Median Income	% Bachelor or more	Funding per Dollar	% Foreign
Overall	0.2495	52.5166	.2749	1.1698	0.0904
Democrat	0.2899	55.8925	.2968	0.9721	0.1210
Republican	0.2122	49.4004	.2547	1.3523	0.0621

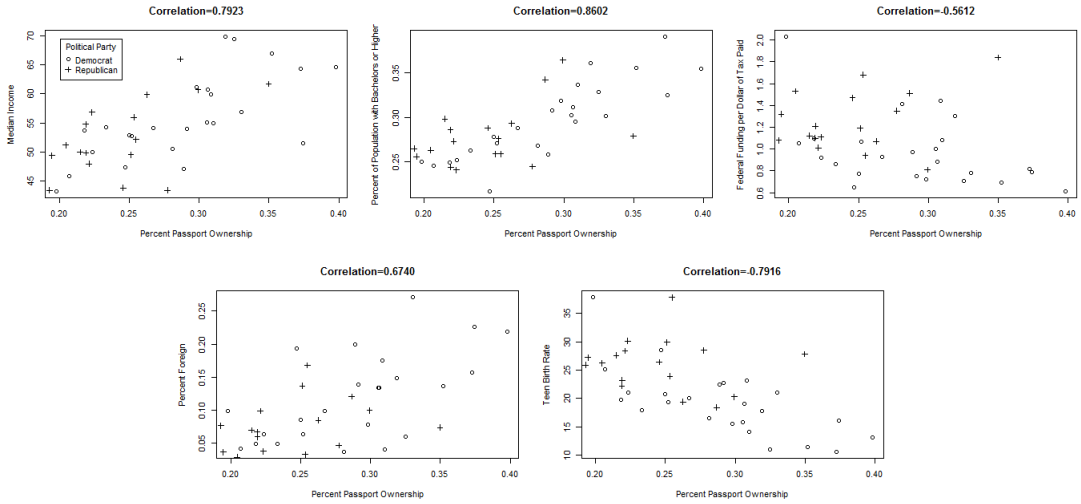

Figure 4.2: Bivariate scatterplots and correlations for the data.

Through the elementary statistical summaries above we see that there are differences among states that vote for either party. The Republican states, as labeled, have lower median incomes, accept more federal funding per dollar of tax paid, have a lower proportion of residents who have received a passport between 2009 and 2015, a lower proportion of residents who are foreign, and a lower proportion of residents

who hold at least a bachelor's degree. Though there are differences in the numerical summaries, Figure 4.2 shows there is much overlap of the two populations.

There are two classes of observations and the feature set is five-dimensional. The goal is to predict whether a state is Republican or Democrat based on the five features described above. In a ten-fold cross validation analysis, the proposed methodology has an error rate of 0.2506 which is slightly lower than the next best classifiers SVM with a radial basis function kernel and Naive Bayes' which tied with an error rate of 0.2607.

4.4.2 Iris Data

The iris dataset is a classic dataset for classification introduced by Fisher (1936). There are three classes of observations and the feature set is four-dimensional. The goal is to predict the species of iris based on its four features: the length and width of the sepals and petals. In a ten-fold cross validation analysis, the proposed methodology has an error rate of 0.0167 which ties both LDA and QDA as the best classifiers; SVM with a linear kernel and KNN were closely behind with a slightly higher error rate of 0.0222.

4.4.3 Titanic Data

A dataset that has enjoyed a recent increase in analyses, largely due to its Kaggle competition and recent updating, is the Titanic survival dataset. There are two classes of observations and the feature set is two-dimensional consisting of the continuous predictors. The goal is to predict whether or not a passenger survived based on two continuous features the age and fare; note there are other qualitative predictors which are left out of this analysis. In a ten-fold cross validation analysis, the proposed methodology has an error rate of 0.316 which was second to SVM with a radial basis function kernel which had a slightly lower error rate of 0.3145.

4.4.4 Classification Maps and Performance Over Varying Sample Sizes

In the following classification example consider the usual bivariate Gaussian density with $\boldsymbol{\mu} = (0\ 0)'$ and covariance matrix $\boldsymbol{\Sigma} = cI_2$ where scalar $c = 5$ and I_2 is the 2×2 identity matrix for class one and for class two a dog-bowl shaped density given by:

$$f(x,y) = (2\pi)^{\frac{-3}{2}} \left(x^2 + y^2\right)^{\frac{-1}{2}} e^{\left(\frac{1}{2}\left(\sqrt{x^2+y^2}-10\right)^2\right)}.$$

A ten-fold cross validation analysis yielded error rates of 0.1091, 0.0746, 0.0462 and .0400 when analyzing this simulated case with samples sizes of $n = 50$, 250, 500, 1000 respectively. As the sample size increases the error rates noticeably decrease which is a direct result of the proposed method's increased ability to estimate the bivariate density as seen in Figure 4.4. We also see that the classification map provides a flexible boundary in Figure 4.3.

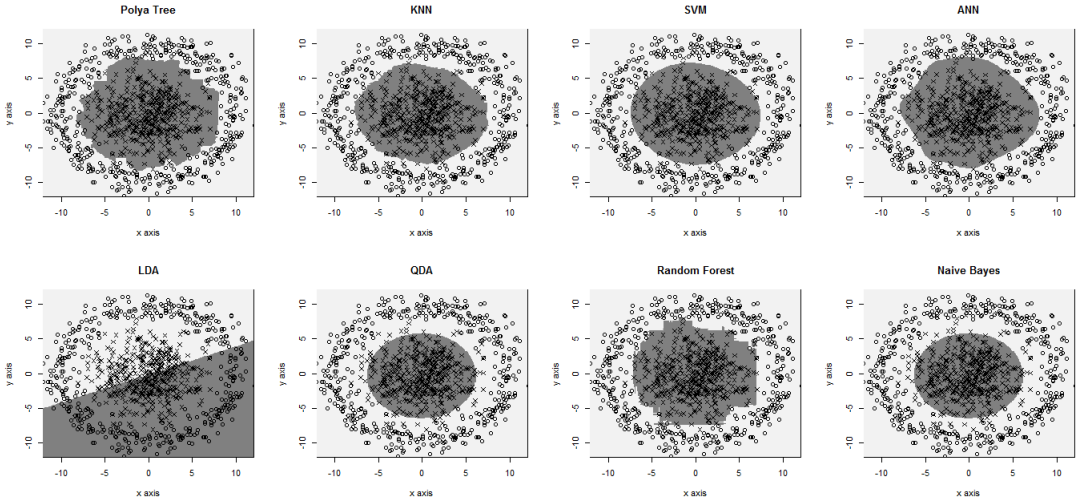

Figure 4.3: Classification maps for various methods for n=500.

4.4.5 Machine Learning Benchmarks

Leisch and Dimitriadou (2015) provide a repository of machine learning benchmark problems in the `mlbench` package for R. Below, in Table 4.2, the results on several artificial machine learning problems are reported using the proposed methodology and several standard approaches. In each simulated case, $n = 400$ data points

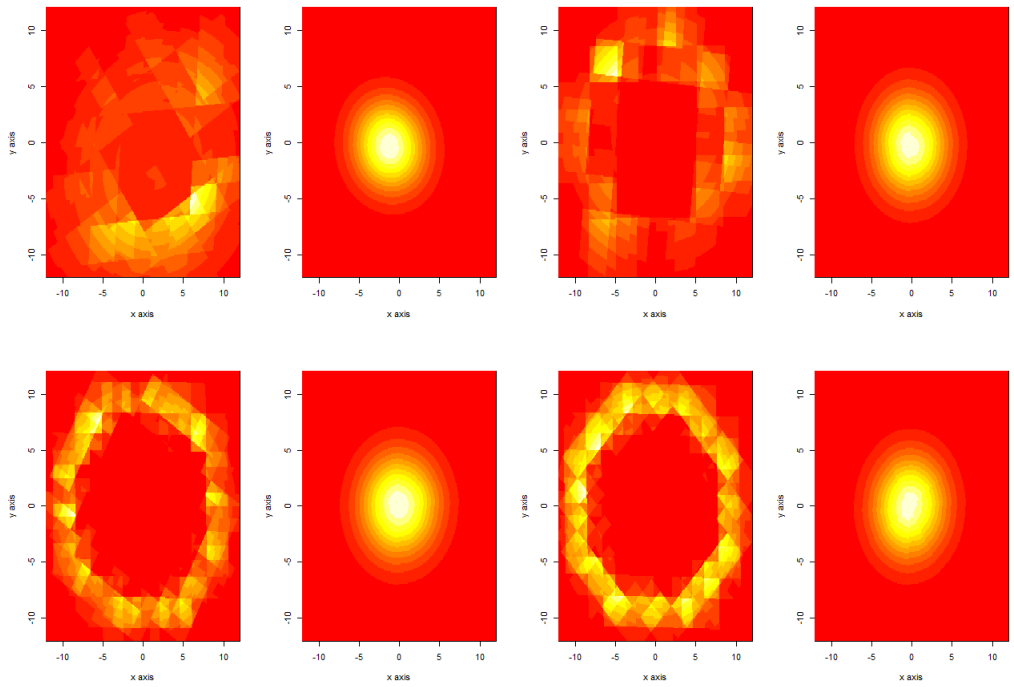

Figure 4.4: Bivariate density estimates for the two classes across samples sizes. Top left, top right, bottom left and bottom right show density estimates for both bivariate distributions for $n = 50, 250, 500$, and 1000 respectively.

are sampled for each class one hundred times and each method is evaluated using the average ten-fold cross validation error rate across the one hundred simulations.

"2dnormals" refers to the case where data consists of two-dimensional Gaussians with unit covariance and g classes, the centers of which are equally spaced on a circle around the origin with radius related to g. Specifically the radius was set to the square root of the number of classes. i.e. data with more classes have a larger radius.

"Threenorm" refers to the case where data consists p-dimensional Gaussians and two classes. The first class is sampled from a multivariate Gaussian with identity covariance and mean vector $(\mu_1, \ldots, \mu_p)$ with each $\mu_i = 2/\sqrt{p}$ and the second class is drawn from a multivariate Gaussian with identity covariance matrix and mean vector $(\mu_1, -\mu_2, \mu_3, \ldots, \mu_p)$ for p odd and $(\mu_1, -\mu_2, \mu_3, \ldots, -\mu_p)$ for p even.

"Circle" refers to the case where data consists of p dimensions and two classes.

The first class is a circle in $p = 2$ dimensions and a p-dimensional sphere for $p > 2$. The second class is a square around the first class in $p = 2$ dimensions and a p-dimensional cube around the first class for $p > 2$. This is considered for a circle, sphere, and hypersphere.

"Spiral" refers to the case where data consists of two dimensions and two classes. The two classes are entangled spirals with added noise to each data point. Here the noise was drawn from a zero-mean Gaussian distributions with standard deviation $\sigma = 0.15$.

"XOR" refers to the case where data consists of two dimensions and two classes. The data make up a square, centered at the origin, with corners at ± 1. The first class is uniformly distributed on the square in the first and third quadrants and the second class is uniformly distributed on the second and fourth quadrants.

"Simplex" refers to the case where data consists of three dimensions and four classes. Each class is sampled from a three-dimensional Gaussian with standard deviation $\sigma = .1$ and means from the corners of a three-dimensional simplex. A simplex is the generalization of a tetrahedral region of space; a simplex in one dimension is a line segment, in two dimensions is the convex hull of an equilateral triangle, and in three dimensions is the convex hull of tetrahedron, a polyhedron with four triangular faces.

"Cuboids" refers to the case where the data consists of three dimensions and four classes. Three classes are uniformly sampled from three, three-dimensional cuboids and the fourth is uniformly distributed on a small cube in the middle of the three cuboids.

"HyperCube" refers to the case where the data consists of three dimensions and six classes. The three classes are sampled from three-dimensional Gaussians with standard deviation $\sigma = .1$ and means from the corners of a cube in three dimensions.

"Cassini" refers to the case where the data consists of two dimensions and three

classes. Two classes are uniformly sampled from two bean shaped areas in two-dimensional space and the third class is uniformly sampled from a circular area between them.

"RingNorm" refers to the case where the data consists of two dimensions and two classes. The first class is sampled from a zero-mean multivariate Gaussian with covariance $4\mathbf{I}$ and the second class is sampled from a zero-mean multivariate Gaussian with identity covariance matrix. In this case, it is expected that the second class is contained within the first class.

"Two Norm" refers to the case where the data consists of p dimensions and two classes. The first class is sampled from a multivariate Gaussian with mean vector $(\mu_1, \ldots, \mu_p)$ with each $\mu_i = 2/\sqrt{p}$ and the second class is drawn from a multivariate Gaussian with unit covariance and mean vector $(-\mu_1, \ldots, -\mu_p)$.

"Waveform" refers to the case where the data consists of twenty one dimensions and three classes. Each class is generated from a combination of two of three "base" waves as in Breiman (1996).

"Shapes" refers to the case where the data consists of two dimensions and four classes. The first class is uniformly distributed across a triangle in the first quadrant; the second class is bivariate Gaussian in the second quadrant; the third class is uniformly distributed across a square in the third quadrant; and the fourth class is uniformly distributed over a wave function in the fourth quadrant.

"Smiley" refers to the case where the data consists of two dimensions and four classes: left eye, right eye, nose and mouth. The first two classes are bivariate Gaussian with standard deviation $\sigma = 0.5$ and make up the eyes of the smiley face; the third class is uniformly distributed over a trapezoid for the nose; and the fourth class is a parabola for the mouth with vertical Gaussian noise with zero-mean and standard deviation $\sigma = 0.5$.

Also considered are the following real-world problems included in this package.

For the real-data analyses a single 10-fold cross validation error rate is provided.

"Breast Cancer" refers to data on $n = 683$ cells from the Wisconsin Breast Cancer Database (Wolberg and Mangasarian, 1990). The nine-dimensional feature set includes observations of clump thickness, uniformity of the cell size, uniformity of the cell shape, marginal adhesion, single Epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses. The goal is to identify new cells as benign, or cancer free, or alternatively malignant, or cancerous.

"Ionosphere" refers to data on $n = 351$ initially analyzed by Sigillito et al. (1989). Each observation is derived from radar data collected in Goose Bay in the Canadian province of Newfoundland and Labrador. The thirty one-dimensional feature set consists of measurements on signals received on sixteen antennas over seventeen pulses. In the original data there were two readings per pulse, meaning thirty-four observed readings, but just thirty-one are considered here as three readings were zero for many of the observations. The goal is to identify which observations are "good," meaning the observations intimate evidence of some structure in the Ionosphere and which are "bad," meaning the observations don't evidence of some structure in the Ionosphere.

"Letter Recognition" refers to data on $n = 20,000$ unique pixel representations of capital letters initially analyzed by Frey and Slate (1991). Each observation is a black and white rectangular display pixel by pixel of one of twenty six capital letters from the English alphabet with added noise to provide unique observations. Each display is summarized by a sixteen-dimensional feature set which includes statistical summaries, i.e. mean, variance, correlation etc. and edge counts both left to right and top to bottom. The goal is to identify the letter of new observations based on the observed feature space.

"Pima Indians" refers to data on $n = 392$ Pima Indians initially analyzed by Wahba et al. (1993). The eight-dimensional feature set includes the number of times a participant is pregnant, as well as seven other measurements on the participant

including age, plasma glucose concentration level, diastolic blood pressure, triceps skin fold thickness, two hour serum insulin, body mass index (BMI), and diabetes pedigree function assessment which represents how likely the participant is to get the diseased based on ancestor's history. The goal is to identify whether or not a new a cell is benign, a cancer free growth, or malignant, a cancerous growth based off of its observations.

"Satellite" refers to data on $n = 6,435$ neighborhoods in a satellite image. Each observation corresponds to a 3×3 square neighborhood of pixels within the satellite image. The thirty six-dimensional feature set is comprised of pixel values in four different spectral bands for each of the nine pixels in each 3×3 square. The goal is to identify the soil type of a 3×3 square via the feature set. The classes are red soil, cotton crop, grey soil, damp grey soil, soil with vegetation stubble and very damp grey soil.

"Sonar" refers to data on $n = 208$ observations of sonar signals bounced off of a metal cylinder or rock as in Gorman and Sejnowski (1988). The sixty-dimensional feature set is comprised of sixty energy measures on different frequency bands over time. The goal is to identify new observations as sonar bounced off metal or sonar bounced off rock.

"Vehicle" refers to data on $n = 846$ observations of vehicle silhouettes initially analyzed by Siebert (1987). The eighteen-dimensional feature set includes various measurements of compactness, circularity, rectangularity, length and statistical summaries, i.e. variance, skewness, kurtosis etc. The goal is to identify the vehicle type of new observations based on the observed feature space. The four vehicle used in this experiment are a Saab 9000 (full size), an Opel Manta 400 (sports car), Chevrolet van (van), and a double decker bus (bus).

"Vowel" refers to data on $n = 990$ observations of the speech coding of eleven steady state vowels of British English. The ten-dimensional feature set includes nine

speech coding measurements, via LPC derived log area ratios. The goal is to identify the vowel type of new observations based on the observed feature space.

Table 4.2 shows the results for all the simulation and data analysis problems as described above. The proposed methodology keeps test error relatively low in the extensive simulation and application here; this is particularly the case when considering classes drawn from d-dimensional Gaussians. The supervised Polya tree approach places in the best three models, among those considered, in terms of error rates for many of the problems described above. When any of the classes are drawn from distributions that strongly deviate from Gaussian, the flexibility of the proposed methodology still does well, often staying within a few percentage points of the best model with two exceptions; both exceptions deal with the "circle" scenario. The first exception is in three dimensions where the first class is a sphere and the second class is a cube with the sphere cut out of the middle. The second exception is in four dimensions where the first class is a hypersphere and the second class is a hypercube with the hypersphere cut out of the middle. The Polya tree centered at the d-dimensional Gaussian has trouble capturing the classes from the three-dimensional cube and four-dimensional hypercube. A simple solution to this problem is to instead use a Polya tree centered at the d-dimensional uniform distribution which would greatly decrease the reported error rate.

Table 4.2: Errors for cross validation classification.

Data	p	g	Polya	KNN	SVMlin	SVMpoly	SVMrad	SVMsig	ANN	LDA	QDA	RF	NB
2dnormals	2	3	0.1162	0.1234	0.1167	0.1333	0.1174	0.1747	0.1187	0.1162	0.1161	0.1317	0.1163
2dnormals	2	6	0.2195	0.2360	0.2197	0.2415	0.2207	0.3352	0.2210	0.2193	0.2197	0.2515	0.2195
threenorm	2	2	0.1342	0.1148	0.1615	0.1704	0.1078	0.2567	0.1124	0.1617	0.1141	0.1213	0.1136
threenorm	3	2	0.1478	0.1269	0.1698	0.1735	0.1204	0.2631	0.1277	0.1697	0.1266	0.1305	0.1433
circle	2	2	0.0592	0.0332	0.4634	0.3893	0.0211	0.4843	0.0195	0.4810	0.0811	0.0331	0.0793
circle	3	2	0.3129	0.1026	0.4686	0.3915	0.0376	0.4465	0.0741	0.4863	0.0944	0.0757	0.0852
circle	4	2	0.3823	0.1539	0.4799	0.4239	0.0420	0.4653	0.1726	0.5132	0.1140	0.1058	0.0961
spiral	2	2	0.4218	0.3494	0.5653	0.4860	0.6676	0.5513	0.6384	0.5690	0.6408	0.4330	0.6359
XOR	2	2	0.0428	0.0303	0.4282	0.4417	0.0275	0.5148	0.0221	0.4916	0.0275	0.0039	0.4911
simplex	3	4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005	0.0000	0.0000	0.0001	0.0000
cuboid	3	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000
hypercube	3	6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0009	0.0000	0.0000	0.0000	0.0000
cassini	2	3	0.0001	0.0000	0.0053	0.0120	0.0000	0.2925	0.0020	0.0040	0.0002	0.0005	0.0002
ringnorm	2	2	0.2427	0.2727	0.3346	0.4024	0.2484	0.4393	0.2511	0.3595	0.2429	0.2834	0.2430
twonorm	2	2	0.0226	0.0242	0.0230	0.0293	0.0232	0.0234	0.0279	0.0226	0.0227	0.0269	0.0227
twonorm	3	2	0.0226	0.0245	0.0233	0.0266	0.0236	0.0231	0.0307	0.0225	0.0227	0.0270	0.0227
twonorm	6	2	0.0227	0.0256	0.0238	0.0253	0.0245	0.0232	0.0359	0.0228	0.0234	0.0282	0.0224
waveform	21	3	0.1714	0.1709	0.1519	0.1584	0.1500	0.1744	0.1881	0.1576	0.1716	0.1549	0.1961
shapes	2	4	0.0001	0.0000	0.0000	0.0005	0.0000	0.0000	0.0001	0.0000	0.0001	0.0000	0.0001
smiley	2	4	0.0797	0.0850	0.1030	0.1666	0.0824	0.1399	0.0790	0.1323	0.0838	0.0744	0.0836
Breast Cancer	9	4	0.0432	0.0318	0.0331	0.0447	0.0331	0.0401	0.0559	0.0418	0.0502	0.0287	0.0387
Letter Recognition	16	4	0.1088	0.0499	0.148	0.103	0.0513	0.5264	0.7946	0.2973	0.1145	0.0322	0.358
Pima Indians	8	4	0.2314	0.2635	0.2293	0.2391	0.239	0.2805	0.3313	0.2219	0.2193	0.2196	0.2314
Satellite	36	4	0.1573	0.1229	0.146	0.1491	0.1171	0.3057	0.635	0.1711	0.1609	0.1013	0.2127
Sonar	60	4	0.3363	0.4449	0.3667	0.3227	0.3307	0.3245	0.3053	0.3931	0.3363	0.2699	0.3831
Vehicle	18	4	0.144	0.3689	0.1954	0.2639	0.2166	0.4855	0.7006	0.2133	0.1452	0.2474	0.5317
Vowel	10	4	0.4436	0.3836	0.4889	0.4085	0.2926	0.6785	0.5267	0.5113	0.4426	0.3241	0.5113
Ionosphere	31	4	0.1222	0.184	0.1792	0.2544	0.0511	0.1862	0.1305	0.1681	0.1195	0.0615	0.1968

4.5 Discussion

The proposed supervised learning procedure using Polya trees is very successful in correctly classifying observations in a variety of scenarios. This approach is computationally efficient and can handle a large number of classes. The flexibility of this model is displayed through the data analyses and simulations of Section 4 where classification in various settings covering a variety of dimensions, sample sizes and number of classes were successful.

The performance of classification was evaluated using ten-fold cross validation which was kept low during simulation and data analysis. A nice aspect of the methodology is that it is computationally efficient and it is capable of providing an approximation of the d-dimensional density even in the case when the density is non-Gaussian. The error rates are still kept relatively low for small sample sizes but, as noted in Section 4, the model does require moderate sample sizes to yield decent density estimates.

Chapter 5

Future Work

5.1 Improved Bayesian Multiple Testing

As seen in Chapter 2 the discrete approximation to a Polya tree prior performed well for multiple testing and estimation of the non-null distribution. A simple improvement on the suggested model in Chapter 2 is to make the Polya tree median-m instead of median-zero. In the motivating example of DNA microarray data, gene expression values tend to move in congress, meaning that the change in expression values across genes is often not median-zero. Previously, we needed to check the data and shift it to be median-zero before analysis; using topics in Chapter 3 we can add m to the list of values to update during MCMC sampling and center the Polya tree at m thus avoiding the need for median-zero data. The information of how the data moves in congress, i.e. the median change from zero, may be of importance to scientists as well.

Going forward, with Chapter 3 in mind, we can attempt to utilize the smoothing techniques introduced to smooth the non-null density estimation which tends to be quite spiky. The ability to better estimate the non-null density is obviously an improvement in its own right, but this improved density estimate should play a key role in the final proposed improvement. The last concern about this approach is that to find "interestingly different" observations one has to provide specific prior information. In many cases trusted expert information is available but it is sensible that scientists may hesitate to trust the results of a model that, in a sense, allows the user

to decide what "interestingly different" observations are through their specification of a prior on w. To improve upon the approach suggested in Chapter 2 we propose another user input ψ which designates a nominal difference in gene expression that a scientist might find "interesting." Through this we will need to change the methodology from testing point mass at zero to testing the interval $A = (m - \psi, m + \psi)$, where units deemed to be an element of this interval are not "interestingly different."

These improvements should make the model more flexible, easier to trust, and lead to better non-null density estimation. The most compelling part of these changes is that we extend the multiple testing for point mass at zero methodology to multiple testing for an interval without the adaptive methodology initially proposed. These compelling improvements will be best displayed when we compare our models multiple testing for an interval to our approximation for the model of Scott and Berger (2006) with the same extension; when testing point mass at zero inaccurately estimating the non-null distribution actually does not always have a very big impact on the ability to keep the combined rate of false discoveries and non-discoveries small. When the problem is extended to testing an interval the non-null density estimation becomes more important.

To see this, consider the simulation where the means follow a skewed, bimodal, median-zero mixture of two Gaussians from Section 2.6.2. Our previously suggested solution does a very good job with multiple testing and density estimation for the non-null distribution when testing point mass at zero. We notice, however, that the approach provided by Scott and Berger (2006) which assumes the means follow a Gaussian distribution does comparably well even though the means follow G_2 which is decidedly not Gaussian, not even close. Scott and Berger (2006) yield a Gaussian non-null density estimate superimposed over the actual bimodal distribution; which has a similar estimate at zero allowing their density estimation miss to not be a malady when testing point mass. We expect it to be a problem when extended to

testing an interval as the density estimate is important at points other than zero. Consider the following illustration: with G_2 the median is between the two modes of the distribution. An accurate density estimation would show that units with means close to the median have a low probability of being from the non-null density. The superimposed Gaussian non-null density estimate of Scott and Berger (2006) would show that units with means close to the median have a high probability of being from the non-null density which should lead to a significant increase in the false discovery rate for that model.

After developing this new and improved model we can revisit the previous analyses in Chapter 2, except this time we don't have to shift the data and we can illustrate the improvement by testing an interval instead of using adaptive methodology to test for "interesting" difference.

5.2 Improved Classification and Prediction

Currently the proposed methodology in Chapter 4 only considers continuous features. A desirable extension would be to handle categorical input via dummy binary variables. Consider the case where each feature vector $\mathbf{x}$ consists of data points $\mathbf{x} = (x_1, \ldots, x_p)'$ containing both continuous and binary data. We can order the observations such that $\mathbf{x} = (x_1, \ldots, x_c, x_{c+1}, \ldots, x_p)' = (\mathbf{x}'_{cont}, \mathbf{x}'_{disc})$ where $\mathbf{x}_{cont} \in \mathbb{R}^c$ and $\mathbf{x}_{disc} \in \{0, 1\}^{p-c}$. We can then consider fitting a model similar to our Polya tree approach in Chapter 4 or we can consider using marginal Polya trees for each binary or discrete input. Another interesting avenue to consider is to consider prediction; the idea is similar to that of Chapter 4, except instead of having a categorical response y to which we classify new observations, there is a quantitative response which we predict for new observations. Lastly, there was one simulation for which our model had a difficult time capturing one of the classes - we intimated that this was due to the choice of centering distribution; a nice extension to the R code would be to allow

the user to identify the centering distribution or to have the choice of seeing output over various centering distributions similar to what is done for SVM kernel selection.

Chapter 6

Conclusion

The suggested approximate finite Polya tree multiple testing procedure is very successful in correctly classifying the observations with non-zero mean in a computationally efficient manner. This holds even when the non-zero means are simulated from a mean zero distribution, as seen in the simulation of θ_i from a $N(0, 2^2)$, which is particularly impressive as we can expect many of these 'non-zero' means to be very close to zero. The flexibility of this model is displayed through the data analyses in Section 7 by completing the task of multiple testing in the cases of proportional differences as well as paired and two sample mean differences. The performance of the multiple comparisons was evaluated using FDR and FNR, both of which were kept low during simulation for relatively small and large numbers of observations. A nice aspect of the methodology and Java applet is that we are able to provide an approximation of the density of the non-zero means that is very close to the actual density function even in the non-Gaussian case. Further, the model is capable of this for "interestingly different" observations in the cases where that is of interest as in Section 6.5 and Section 7.1. This model assumes that θ_i and σ_i^2 are independent and is sensitive to the prior specifications including w. The model is also sensitive to the data being median zero; simulation results, not included, show that deviations from this lead to inflated error. A discrete approximation to the univariate MPT model is proposed that has conjugate and efficient MCMC updating for all parameters. A smoothed version, the SAPT, maintains conjugate updating and mixes significantly better than the traditional MPT. Various simulated and real data illustrate the use-

fulness of the proposed approach. Multivariate versions of the APT and SAPT models suffer the same fate as multivariate extensions of B-splines: an exponential explosion of parameters occurs with increased dimensionality. For example the nonparametric random effects distribution of Ghidey et al. (2004) is largely relegated to bivariate densities. Finally, the proposed supervised learning procedure using Polya trees is very successful in correctly classifying observations in a variety of scenarios. This approach is computationally efficient and can handle a large number of classes. The flexibility of this model is displayed through the data analyses and simulations of Section 4 as we complete classification in various settings covering a variety of dimension, sample sizes and number of classes. The performance of classification was evaluated using ten-fold cross validation which was kept low during simulation and data analysis. A nice aspect of the methodology is that it is computationally efficient and it is capable of providing an approximation of the d dimensional density even in the case when the density is non-Gaussian. The error rates are still kept relatively low for small sample sizes but, as noted in Section 4, the model does require moderate sample sizes to yield decent density estimates.

Bibliography

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons.

Aitchison, J. and Shen, S. M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika*, 67:261–272.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55:117–128.

Aldous, D. (1985). Exchangeability and related topics. In *Ecole d'Ete de Probabilities de Saint-Flour XIII 1983*, pages 1–198. Springer.

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750.

Alpaydin, E. (2014). *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Anderson, J. A. and Rosenfeld, E., editors (1988). *Neurocomputing: Foundations of Research*. MIT Press, Cambridge, MA, USA.

Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2:1151–1174.

Atiya, A. F., Fayed, H. A., and Abdel-Gawad, A. H. (2013). A new Monte Carlo based algorithm for the Gaussian process classification problem. *ArXiv e-prints*.

Ausin, M., Gomez-Villegas, M., Gonzalez-Perez, B., Rodriguez-Bernal, M., Salazer, I., and Sanz, L. (2011). Bayesian analysis of multiple hypothesis testing with applications to microarray experiments. *Communications in Statistics - Theory and Methods*, 40(13):2276–2291.

Bajgrowicz, P. and Scaillet, O. (2007). Technical trading revisited: False discoveries, persistence tests, and transaction costs. Technical report.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Soceity. Series B*, 57(1):289–300.

Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D., and Li, S. (2013). *FNN: Fast Nearest Neighbor Search Algorithms and Applications*. R package version 1.1.

Blackwell, D. (1973). Discreteness of Ferguson selections. *The Annals of Statistics*, 1:356–358.

Blei, D., Griffiths, T., Jordan, M., and Tenenbaum, J. (2004). Hierarchical topic models and the nested Chinese restaurant process. *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference*.

Blei, D. M. and Frazier, P. I. (2011). Distance dependent Chinese restaurant processes. *Journal of Machine Learning Research*, 12:2461–2488.

Blei, D. M., Griffiths, T., and Jordan, M. (2010). The nested Chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57:7:1–7:30.

Blomquist, J. (2014). Multiple inference and market integration: An application to Swedish fish markets. *Journal of Agricultural Economics*, 66(1):221–234.

Bogdan, M., Ghosh, J. K., and Tokdar, S. T. (2008). *A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing*, volume 1, pages 211–230. Institute of Mathematical Statistics.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152. ACM.

Branscum, A. and Hanson, T. (2008). Bayesian nonparametric meta-analysis using Polya tree mixture models. *Biometrics*, 64:825–833.

Breiman, L. (1996). Bias, variance, and arcing classifiers. Technical report.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Buckley, J. J. and James, I. R. (1979). Linear regression with censored data. *Biometrika*, 66:429–436.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167.

Burr, D. and Doss, H. (2005). A Bayesian semiparametric model for random-effects meta-analysis. *Journal of the American Statistical Association*, 100:242–251.

Canale, A. and Dunson, D. (2011). Bayesian kernel mixtures for counts. *Journal of the American Statistical Association*, 106:1528–1539.

Canale, A. and Dunson, D. B. (2016). Multiscale Bernstein polynomials for densities. *Statistica Sinica*, in press.

Chen, Y., Hanson, T., and Zhang, J. (2014). Accelerated hazards model based on parametric families generalized with Bernstein polynomials. *Biometrics*, 70:192–201.

Christensen, R., Johnson, W., Branscum, A., and Hanson, T. (2010). *Bayesian Ideas and Data Analysis An Introduction for Scientists and Statisticians*. CRC Press.

Chumbley, J. and Friston, K. (2009). False discovery rate revisited: FDR and topological inference using Gaussian random fields. *NeuroImage*, 44(11):62–70.

Cipolli, W. and Hanson, T. (2016). Computationally tractable approximate and smoothed Polya trees. In revision with Statistics and Computing.

Cipolli, W., Hanson, T., and McLain, A. (2016). Bayesian nonparametric multiple testing. *Computational Statistics and Data Analysis*, 101:64–79.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27.

Craddock, T., Harvey, J., Nathanson, L., Barnes, Z., Klimas, N., Fletcher, M., and Broderick, G. (2015). Using gene expression signatures to identify novel treatment strategies in Gulf War illness. *BMC Medical Genomics*, 8(1).

Dahl, D. B., Kim, S., and Vannucci, M. (2009). Spiked Dirichlet process prior for Bayesian multiple hypothesis testing in random effects models. *Bayesian Analysis*, 4(4):707–732.

Daumé III, H. and Marcu, D. (2005). A Bayesian model for supervised clustering with the Dirichlet process prior. *Journal of Machine Learning Research (JMLR)*, 6:1551–1577.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society,Series B*, 39:1–38.

Do, K., Müller, P., and Tang, F. (2005). A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society*, 54(03):627–644.

Draper, D. (1999). Discussion of Bayesian nonparametric inference for random distributions and related functions. *Journal of the Royal Statistical Society Series B*, 61:510–513.

Duan, K. and Keerthi, S. S. (2005). Which is the best multiclass svm method? an empirical study. In *Proceedings of the Sixth International Workshop on Multiple Classifier Systems*, pages 278–285.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification (2Nd Edition)*. Wiley-Interscience.

Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, 6:325–327.

Efron, B., Tibshirani, R., Storey, J., and Tusher, V. (2001). Empirical Bayes analysis of microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, 02(04):615–629.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188.

Florida, R. (2011). America's great passport divide. [Online; posted 15-Mar-2011].

Follman, D. A. and Lambert, D. (1989). Generalizing logistic regression by nonparametric mixing. *Journal of the American Statistical Association*, 84:295–300.

Frey, P. W. and Slate, D. J. (1991). Letter recognition using Holland-style adaptive classifiers. *Machine Learning*, 6:161–182.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*, 7:57–68.

Gans, P. and Gill, J. (1984). Smoothing and differentiation of spectroscopic curves using spline functions. *Applied Spectroscopy*, 38:370–376.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. 74:153–160.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B*, 56:501–514.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery procedure. *Journal of the Royal Statistical Soceity. Series B*, 64(3):499–517.

Ghidey, W., Lesaffre, E., and Eilers, P. (2004). Smooth random effects distribution in a linear mixed model. *Biometrics*, 60:945–953.

Ghosh, A. (2006). On optimum choice of k in nearest neighbor classification. *Computational Statistics and Data Analysis*, 50:3113–3123.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer-Verlag: New York.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., and Bloomfield, C. D. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537.

Gorman, P. R. and Sejnowski, T. J. (1988). Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks*, 1:75–89.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732.

Guindani, M., Müller, P., and Zhang, S. (2009). A Bayesian discovery procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):905–925.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242.

Hannah, L. A., Blei, D. M., and Powell, W. B. (2011). Dirichlet process mixtures of generalized linear models. *J. Mach. Learn. Res.*, 12:1923–1953.

Hanson, T. (2006). Inference for mixtures of finite Polya tree models. *Journal of the American Statistical Association*, 101(0476):1548–1565.

Hanson, T., Branscum, A., and Gardner, I. (2008). Multivariate mixtures of Polya trees for modelling ROC data. *Statistical Modelling*, 8:81–96.

Hanson, T. and Jara, A. (2013). Surviving fully Bayesian nonparametric regression models. *Bayesian Theory and Applications*, pages 593–615.

Hanson, T. and Johnson, W. (2002). Modeling regression error with a mixture of polya trees. *Journal of the American Statistical Association*, 97:1020–1033.

Hanson, T., Monteiro, J., and Jara, A. (2011). The Polya tree sampler: Towards efficient and automatic independent Metropolis-Hastings proposals. *Journal of Computational and Graphical Statistics*, 20:41–62.

Harrell, Jr, F. (2015). *rms: Regression Modeling Strategies*. R package version 4.4-0.

Hassanat, A. B., Abbadi, M. A., and Altarawneh, G. A. (2014). Solving the problem of the k parameter in the KNN classifier using ensemble learning approach. *International Journal of Computer Science and Information Security*, 12:33–39.

Hastie, T. and Tibshirani, R. (1998). Classification by pairwise coupling. *The Annals of Statistics*, 26:451–471.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Higdon, D. (2002). Space and space-time modeling using process convolutions. In Anderson, C., Barnett, V., Chatwin, P., and El-Shaarawi, A., editors, *Quantitative Methods for Current Environmental Issues*, pages 37–56. Springer London.

Hjort, N., Holmes, C., Müller, P., and Walker, S. G., editors (2010). *Bayesian Nonparametrics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press: Cambridge.

Ho, T. K. (1995). Random decision forests. In *Third International Conference on Document Analysis and Recognition, ICDAR 1995, August 14 - 15, 1995, Montreal, Canada. Volume I*, pages 278–282.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance comparing individual means in the analysis of variance. *Biometrika*, 75(4):800–802.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer-Verlag.

Ishwaran, H. and Zarepour, M. (2002). Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 30:269–283.

Izenman, A. J. (1991). Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, 86:205–224.

Jara, A., Hanson, T., and Lesaffre, E. (2009). Robustifying generalized linear mixed models using a new class of mixtures of multivariate Polya trees. *Journal of Computational and Graphical Statistics*, 20:41–62.

Jara, A., Hanson, T., Quintana, F., Müeller, P., and Rosner, G. (2011). Dppackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software*, 40:1–30. http://www.jstatsoft.org/v40/i05/.

Jiang, L., Wang, D., Cai, Z., and Yan, X. (2007). Survey of improving naive Bayes for classification. In *Proceedings of the 3rd International Conference on Advanced Data Mining and Applications*, pages 134–145. Springer-Verlag.

Jin, J. (2008). Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *Journal of the Royal Statistical Society*, 70(3):461–493.

Jin, J. and Cai, T. T. (2007). Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association*, 102(478):495–506.

Karsoliya, S. (2012). Approximating number of hidden layer neurons in multiple hidden layer BPNN architecture. *International Journal of Engineering Trends and Technology*, 12:714–717.

Kleinman, K. P. and Ibrahim, J. G. (1998). A semi-parametric Bayesian approach to generalized linear mixed models. *Statistics in Medicine*, 17:2579–2596.

Komárek, A. and Lesaffre, E. (2008). Generalized linear mixed model with a penalized Gaussian mixture as a random-effects distribution. *Computational Statistics & Data Analysis*, 52:3441–3458.

Komárek, A., Lesaffre, E., and Hilton, J. (2005). Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *Journal of Computational and Graphical Statistics*, 14:726–745.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification. *Informatica*, 31:249–268.

Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armañanzas, R., Santafé, G., and Pérez, A. (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, 17:86–112.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics*, 20(03):1222–1235.

Lavine, M. (1994). More aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics*, 22(03):1161–1176.

Ledl, T. (2004). Kernel density estimation: Theory and application in discriminant analysis. *Austrian Journal of Statistics*, 33:267–279.

Leisch, F. and Dimitriadou, E. (2015). *mlbench: Machine Learning Benchmark Problems*. R package version 2.1-1.

Liaw, A. and Wiener, M. (2002). Classification and regression by random forest. *R News*, 2:18–22.

Liu, L., Lei, J., Sanders, S. J., Willsey, A. J., Kou, Y., Cicek, A. E., Klei, L., Lu, C., He, X., Li, M., Muhleand, R. A., MaâĂŹayan, A., Noonan, J. P., Sestan, N., McFadden, K. A., State, M. W., Buxbaum, J. D., Devlin, B., and Roeder, K.

(2014). Dawn: a framework to identify autism genes and subnetworks using gene expression and genetics. *Molecular Autism*, 5(1).

Longnecker, M. (1994). Alcoholic beverage consumption in relation to risk of breast cancer: Meta-analysis and review. *Cancer Causes and Control*, 5:73–82.

Ma, Y. and Guo, G. (2014). *Support Vector Machines Applications*. Springer International Publishing.

Martin, R. and Tokdar, S. (2012). A nonparametric empirical Bayes framework for large-scale multiple testing. *Biostatistics*, 13(3):427–439.

Marzio, M. and Taylor, C. C. (2005). On boosting kernel density methods for multivariate data: density estimation and classification. *Statistical Methods and Applications*, 14:163–178.

Mauldin, R. D., Sudderth, W. D., and Williams, S. C. (1992). Polya trees and random distributions. *The Annals of Statistics*, 20(03):1203–1221.

Mcculloch, W. S. and Pitts, W. H. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133.

McMillan, G. (2001). *Ache Residential Grouping and Social Foraging*. PhD thesis, University of New Mexico.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2015). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.6-7.

Migration Policy Institute (2014). State immigration data profiles. [Online; accessed 13-Mar-2016].

Minka, T. P. (2001). *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Cambridge, MA, USA.

Minsky, M. and Papert, S. (1969). *Perceptrons; an Introduction to Computational Geometry*. MIT Press.

Mitra, R. and Müller, P., editors (2015). *Nonparametric Bayesian inference in Biostatistics*. Frontiers in Probability and the Statistical Sciences. Springer International Publishing: Switerland.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning*. The MIT Press.

Morley, M., Molony, C., Weber, T., Devlin, J., Ewens, K., and Spielman, R. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature*, 430:742–747.

Mukhopadhyay, S. and Ghosh, A. (2011). Bayesian multiscale smoothing in supervised and semi-supervised kernel discriminant analysis. *Computational Statistics and Data Analysis*, 55:2344–2353.

Müller, P., Parmigiani, G., and Rice, K. (2007). FDR and Bayesian multiple comparisons rules.

Müller, P., Quintana, F., Jara, A., and Hanson, T. (2015). *Bayesian Nonparametric Data Analysis*. Springer: Switzerland.

Müller, P. and Rodriguez, A. (2013). *Chapter 4: Pólya Trees*, volume Volume 9 of *NSF-CBMS Regional Conference Series in Probability and Statistics*, pages 43–51. Institute of Mathematical Statistics and American Statistical Assocation.

Muralidharan, O. (2010). An empirical Bayes mixture method for effect size and false discovery rate estimation. *Journal of the American Statistical Association*, 4(1):422–438.

National Archives and Records Administration (2012). Historical election results. [Online; accessed 13-Mar-2016].

Nickisch, H. and Rasmussen, C. (2008). Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*, 9:2035–2078.

Notterman, D. A., Along, U., and Sierk, A. J. (2001). Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Research*, 61(7):3124–3130.

Paddock, S., Ruggeri, F., Lavine, M., and West, M. (2003). Randomised Polya tree models for nonparametric Bayesian inference. *Statistica Sinica*, 13:443–460.

Patti, M. E., Butte, A. J., Crunkhorn, S., Cusi, K., Berria, R., Kashyap, S., Miyazaki, Y., Kohane, I., Costello, M., Saccone, R., Landaker, E. J., Goldfine, A. B., Mun, E., DeFronzo, R., Finlayson, J., Kahn, C. R., and Mandarino, L. J. (2003). Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: Potential role of pgc1 and nrf1. *Proceedings of the National Academy of Sciences*, 100(14):8466–8471.

Peña, E. A., Habiger, J. D., and Wu, W. (2011). Power-enhanced multiple decision functions controlling family-wise error and false discovery rates. *The Annals of Statistics*, 39(1):556–583.

Pitman, J. (2002). Combinatorial stochastic processes. Technical report.

Qin, Z. S. (2006). Clustering microarray gene expression data using weighted Chinese restaurant process. *Bioinformatics*, 22:1988–1997.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society - Series B*, 10:159–203.

Rasmussen, C. E. (1999). The infinite Gaussian mixture model. pages 554–560. The MIT Press.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Rish, I. (2001). An empirical study of the naive Bayes classifier. Technical report, IBM.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in galaxies. *Journal of the American Statistical Association*, 85:617–624.

Rojas, R. (1996). *Neural Networks: A Systematic Introduction*. Springer-Verlag New York, Inc., New York, NY, USA.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Physchological Review*, 65:386–408.

Sargent, D. J., Hodges, J. S., and Carlin, B. P. (2000). Structured Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 9:217–234.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

Scholkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA.

Scott, J. G. and Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136(07):2144–2162.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.

Shahbaba, B. and Neal, R. (2009). Nonlinear models using Dirichlet process mixtures. *J. Mach. Learn. Res.*, 10:1829–1850.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633.

Siebert, J. (1987). *Vehicle Recognition Using Rule Based Methods*. TIRM–87-018. Turing Institute.

Sigillito, V. G., Wing, S. P., Hutton, L. V., and Baker, K. B. (1989). Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest*, pages 262–266.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer International Publishing.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Soceity. Series B*, 64(3):479–498.

Sun, W. and McLain, A. C. (2012). Multiple testing of composite null hypotheses in heteroscedastic models. *Journal of the American Statistical Association*, 107(398):673–687.

Tax Foundation (2007). Federal taxes paid vs. federal spending received by state, 1981-2005. [Online; accessed 13-Mar-2016].

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2005). Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes. *Advances in Neural Information Processing Systems 17*, pages 1385–1392.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 5(2):99–114.

United States Census Bureau (2010). American community survey, education attainment for states, percent with high school diploma and with bachelor's degree: 2010. [Online; accessed 13-Mar-2016].

United States Census Bureau (2014). State median income. [Online; accessed 13-Mar-2016].

United States Department of State Bureau of Consular Affairs (2015). U.S. passports and international travel: Passport statistics. [Online; accessed 13-Mar-2016].

Unser, M., Aldroubi, A., and Eden, M. (1992). On the asymptotic convergence of B-spline wavelets to Gabor functions. *IEEE Transactions on Information Theory*, 38:864–872.

van 't Veer, L., Dai, H., van de Vijver, M., He, Y., Hart, A., Mao, M., Peterse, H., van der Kooy, K., Marton, M., Witteveen, A., Schreiber, G., Kerkhoven, R., Roberts, C., Linsley, P., Bernards, R., and Friend, S. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536.

Vapnik, V. N. (1979). *Estimation of Dependences Based on Empirical Data [in Russian]*. Nauka, USSR.

Vapnik, V. N. and Chervonenkis, A. (1963). A note on one class of perceptrons. *Automation and Remote Control*, 25:774–780.

Vapnik, V. N. and Lerner, A. (1962). Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:709–715.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Wahba, G., Gu, C., Wang, Y., and Chappell, R. (1993). Soft classification, a. k. a. risk estimation, via penalized log likelihood and smoothing spline analysis of variance. In *The Mathematics of Generalization*. Addison-Wesley.

Werbos, P. J. (1981). Applications of advances in nonlinear sensitivity analysis. In *Proceedings of the 10th IFIP Conference, 31.8 - 4.9, NYC*, pages 762–770.

Widrow, B. (1962). Generalization and information storage in networks of adaline 'neurons'. In Yovits, M. C., Jacobi, G. T., and Goldstein, G. D., editors, *Self-Organizing Systems 1962*. Sparton, Washington.

Widrow, B. and Hoff, M. E. (1960). Adaptive switching circuits. *Institute of Radio Engineers, Western Electronics Show and Convention*, Part 4:96–104.

Williams, C. K. I. and Barber, D. (1998). Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:1342–1351.

Wolberg, W. H. and Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences,U.S.A.*, 87:9193–9196.

Wong, W. H. and Ma, L. (2010). Optional Polya tree and Bayesian inference. *The Annals of Statistics*, 38:1433–1459.

Wu, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912.

Yegnanarayana, B. (2004). *Artificial Neural Networks*. Prentice-Hall of India Pvt.Ltd.

Zambom, A. Z. and Dias, R. (2013). A review of kernel density estimation with applications to econometrics. *International Econometric Review (IER)*, 5:20–42.

Zhao, L. and Hanson, T. (2011). Spatially dependent Polya tree modeling for survival data. *Biometrics*, 67:391–403.